UNIVERSITY OF CALIFORNIA SAN DIEGO

Listening to Ice and Ocean: Machine Learning for Seismic and Acoustic Environmental Characterization

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Oceanography

by

William Frost Jenkins II

Committee in charge:

Professor Peter Gerstoft, Chair
Peter Bromirski
Matthew Dzieciuch
Professor William Hodgkiss
Professor Piya Pal

2023

The dissertation of William Frost Jenkins II is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

To my family,

especially

芙美子

EPIGRAPH

The sea, once it casts its spell, holds one in its net of wonder forever.

Jacques Cousteau

the rough sea—	荒海や
stretching above Sado Island	差渡によこたふ
the Milky Way	天の川

Bashō Matsuo 松尾 芭蕉

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would first like to thank my advisor and committee chair, Peter Gerstoft, for his generous support, time, and guidance. His door was always open, and I enjoyed our many fruitful discussions about research and other endeavors. I am truly grateful for the many opportunities I was afforded as a member of Noiselab to explore my interests and grow my professional relationships.

I am grateful to my dissertation committee for their guidance and support of my research. Their feedback and discussions imbued me with new perspectives and insights and improved the quality of my work. I would especially like to thank Professor Hodgkiss for his mentorship and instruction on signal processing. His wisdom and experience were critical to helping me maintain a clear vision for both my research and professional goals.

I would like to thank my coauthors, especially Michael Bianco, Peter Bromirski, and Yongsung Park, whose patience and assistance I deeply appreciated during the preparation of our publications.

I would like to acknowledge my colleagues in Noiselab, who have been a source of continual inspiration, energy, knowledge, and joy.

I would like to acknowledge the many friends and colleagues who have supported and encouraged me in my journey. I would especially like to express my gratitude to Michael Bianco, Yongsung Park, Gihoon Byun, and Hunter Akins for our deep and probing discussions related to research and life in general.

Finally, I would like to thank my family for giving me the strength and courage to pursue my doctoral education. In particular, I am profoundly grateful to my wife, Fumiko Naka Jenkins, for her unwavering love, support, and encouragement across oceans and continents.

Chapter 2, in full, is a reprint of the material as it appears in the Journal of Geophysical Research: Solid Earth 2021. Jenkins, William; Gerstoft, Peter; Bianco, Michael; Bromirski, Peter, American Geophysical Union, 2021. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in the Proceedings of Meetings on Acoustics 2022. Jenkins, William; Johnson, Hayden; Vakhutinsky, Sofia; Helmberger, Meghan; Storheim, Espen; Sagen, Hanne; Sandven, Stein, Acoustical Society of America, 2022. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is a reprint of the material as it appears in the Journal of the Acoustical Society of America 2023. Jenkins, William; Gerstoft, Peter; Park, Yongsung, Acoustical Society of America, 2023. The dissertation author was the primary investigator and author of this paper.

Chapter 5, in full, has been submitted for publication of the material as it may appear in the Proceedings of the Institute of Electrical and Electronics Engineers International Conference on Acoustics, Speech, and Signal Processing 2024. Jenkins, William; Gerstoft, Peter, Institute of Electrical and Electronics Engineers, 2023. The dissertation author was the primary investigator and author of this paper.

Chapter 6, in part, is currently being prepared for submission for publication of the material. Jenkins, William; Gerstoft, Peter; Park, Yongsung. The dissertation author was the primary investigator and author of this material.

VITA

2009	B.S. with Merit in oceanography, United States Naval Academy
2010	M.S. with Distinction in engineering acoustics, Naval Postgraduate School
2009–2017	Submarine Warfare Officer, United States Navy
2017–2023	Submarine Warfare Officer, United States Navy Reserve
2017–2023	Graduate Student Researcher, University of California San Diego
2018–2022	National Defense Science and Engineering Graduate Fellow

PUBLICATIONS

W.F. Jenkins II, P. Gerstoft, Y. Park, "Geoacoustic inversion with Bayesian optimization," *Journal of the Acoustical Society of America*, 2023 (submitted).

W.F. Jenkins II, P. Gerstoft, "Bayesian optimization with Gaussian processes for robust localization," *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2023 (submitted).

W.F. Jenkins II, P. Gerstoft, Y. Park, "Bayesian optimization with Gaussian process surrogate model for source localization," *Journal of the Acoustical Society of America*, 154(3), pp. 1459-1470, 2023, DOI: https://doi.org/10.1121/10.0020839.

C.-C. Chien, **W. Jenkins**, P. Gerstoft, M. Zumberge, and R. Mellors, "Automatic classification with an autoencoder of seismic signals on a distributed acoustic sensing cable," *Computers and Geotechnics*, 155, 2023, DOI: https://doi.org/10.1016/j.compgeo.2022.105223.

W. Jenkins, H. Johnson, S. Vakhutinsky, M. Helmberger, E. Storheim, H. Sagen, and S. Sandven, "Analysis of underwater acoustic data collected under sea ice during the Useful Arctic Knowledge 2021 cruise," *Proc. of Meetings on Acoustics*, 47(1), 2022, DOI: https://doi.org/10.1121/2.0001574.

W.F. Jenkins II, P. Gerstoft, M.J. Bianco, and P.D. Bromirski, "Unsupervised Deep Clustering of Seismic Data: Monitoring the Ross Ice Shelf, Antarctica," *Journal of Geophysical Research: Solid Earth*, 126(9), 2021, DOI: https://doi.org/10.1029/2021JB021716.

W. Jenkins, J. Rice, L. Ziomek, and D. Green, "Multi-channel MFSK Modulation and Demodulation through Short-Range, Large-Bandwidth, Underwater Acoustic Channels," *Proc. of the 10th European Conference on Underwater Acoustics*, 2010.

ABSTRACT OF THE DISSERTATION

Listening to Ice and Ocean: Machine Learning for Seismic and Acoustic Environmental Characterization

by

William Frost Jenkins II

Doctor of Philosophy in Oceanography

University of California San Diego, 2023

Professor Peter Gerstoft, Chair

Seismology and ocean acoustics are important remote sensing tools, enabling observation of environments that are difficult to access and directly measure. Seismic and ocean acoustic remote sensing are data-intensive tasks, and the proliferation of remote sensing systems has led to the generation of vast amounts of data. Meanwhile, advances in machine learning (ML) techniques and computational capacity have yielded state-of-the-art methodologies for processing and analyzing large seismic and acoustic data sets. This dissertation presents two ML-based paradigms for the characterization of environments using seismic and acoustic data.

First, unsupervised ML is demonstrated for automatically identifying dominant types

of seismicity present data recorded from a 34-station broadband seismic array deployed on the Ross Ice Shelf (RIS), Antarctica from 2014 to 2017. The data set contains signals generated by glaciological processes that have been used to monitor the integrity and dynamics of ice shelves. Deep clustering automatically groups these signals into classes without the need for manual labeling, enabling comparison of potential source mechanisms with not only the spatial and temporal distributions of the signals but also their characteristics. The method learns the salient features of spectrograms and encodes them into a lower-dimensional latent representation using an autoencoder, a type of deep neural network. Two clustering methods are applied to the latent data and compared: a Gaussian mixture model (GMM) and deep-embedded clustering (DEC). Dominant types of seismic signals are identified and compared with environmental data such as temperature, wind speed, tides, and sea ice concentration. The highest seismicity occurred at the RIS front during the 2016 El Niño summer, and diurnally near grounding zones throughout the deployment.

The second paradigm presents Bayesian optimization (BO) as a method for efficiently estimating geoacoustic parameters within a fixed computational budget. An objective function is defined using the Bartlett processor, whose output measures the match between a received and predicted pressure field on a vertical line array. BO is a sequential framework that iteratively fits a Gaussian process surrogate model to the objective function and then uses a heuristic acquisition function to select the next point to evaluate. After each evaluation, the GP surrogate model is re-fit, and the optimization proceeds until the budget is expended. BO is demonstrated using both simulations and real data collected during an ocean acoustics experiment. Results indicate BO rapidly estimates the correct parameters and achieves better correlations between observed and predicted data.

Chapter 1

Introduction

Seismology and ocean acoustics have revealed insights and driven understanding of the interior structures and processes of the earth [1, 2], ocean [3, 4, 5, 6], and cryosphere [7, 8]. Seismic and acoustic waves are inextricably shaped by the environments and media through which they propagate. Signal processing, combined with physical theory and intuition made possible by computational models, enables the recovery of information and properties about the environments and media through which propagation occurs. Seismic and acoustic remote sensing are therefore important tools for characterizing environments, particularly those that are difficult to access and measure, such as polar regions, oceans, and seabeds.

Seismic and acoustic sensing are inherently data-intensive tasks. Arrays may contain dozens or hundreds of channels, recording continuously for hours, months, or years at high sampling rates. As these data sets grow larger and more prevalent, labor-intensive, manual analyses performed with conventional signal processing techniques are becoming inadequate for timely and comprehensive analyses of data sets. Furthermore, an increasing demand for autonomous systems to venture into challenging and dynamic environments, such as navigating and communicating beneath sea ice, has led to a need for automated environmental characterization methods. Meanwhile, advances in computing capabilities and machine learning (ML) algorithms have enabled more efficient, data-driven approaches for studying natural processes and phenomena [9, 10, 11, 12, 13, 14, 15].

The objective of this dissertation is to present two ML-based environmental characterization paradigms using seismic and underwater acoustic data:

1. *Unsupervised ML for pattern discovery* relies on clustering of data to assess the frequency of occurrence of certain types of signals. This dissertation specifically explores the association of certain kinds of seismicity with potential environmental source mechanisms [16], but clustering has numerous other applications, such as identifying and classifying unlabeled biological sounds [17]. The technique presented here involves *deep clustering*, which reduces the dimensionality of the input data with a neural network to improve clustering algorithm performance [18, 19, 20, 21, 17]. This paradigm is best suited for the exploration of large data sets, where the priority is to identify dominant or anomalous patterns within the data for further investigation.

2. *Geoacoustic inversion* seeks to estimate the environmental parameters that explain observed acoustic data. Whereas geoacoustic inversion is generally a computationally expensive endeavor requiring thousands of simulations to estimate parameters [22, 6], *Bayesian optimization* [23, 24, 25, 26] is a global optimization framework that attempts to find the optimal parameters as efficiently as possible. To demonstrate the viability and characteristics of Bayesian optimization, two acoustic source localization parameterizations are demonstrated [27, 28] before a more challenging, higher-dimensional optimization is demonstrated for both source localization and geoacoustic inversion.

1.1 Basic concepts

1.1.1 Clustering

Unsupervised machine learning involves algorithms and models that learn patterns and structures from data without explicit supervision or labeled target outputs [14, 15]. Specifically, clustering algorithms seek to discover similar examples within the data [14, 15] and are useful for data mining and exploratory data analysis. While there is a diverse set of paradigms for clustering

algorithms [15, ch. 21], Chapter 2 makes use of algorithms that utilize the distances between data examples to determine similarity and groupings. Consider a data set $\mathscr{D} = \{\boldsymbol{x}_n : n = 1 : N\}$, where $\boldsymbol{x}_n$ is the n^{th} vector representation of a D-dimensional data sample $\boldsymbol{x} = [x_1, \ldots, x_D]^\mathsf{T} \in \mathbb{R}^D$ and superscript T denotes the transpose operator. The Euclidean distance between any two points in $\mathscr{D}$ is given by:

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2. \tag{1.1}$$

One of the most widely used clustering methods is the k-means algorithm [29, 30, 31], which partitions the data $\mathscr{D}$ into K sets $\mathbf{S} = \{S_k : k = 1 : K\}$, each described by a prototypical example of the data assigned to the cluster. The following derivation closely follows that of [14, sec. 9.1]. The prototype $\boldsymbol{\mu}_k \in \mathbb{R}^D$ is taken as the mean of all data assigned to the cluster; in a geometric sense, it is the center, or centroid, of the cluster:

$$\boldsymbol{\mu}_k = \frac{1}{|S_k|} \sum_{\boldsymbol{x} \in S_k} \boldsymbol{x}. \tag{1.2}$$

The goal of k-means clustering is to optimize the assignment of data to clusters and to find the set of centroids $\{\boldsymbol{\mu}_k\}$ that minimize the sum of the squares of the distances of each data point to its closest centroid $\boldsymbol{\mu}_k$. This is accomplished through the minimization of a cost function called the *distortion*:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2, \tag{1.3}$$

where $r_{nk} \in \{0, 1\}$ is a binary variable indicating whether a data point $\boldsymbol{x}_n$ is assigned to cluster k according to:

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\boldsymbol{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases} \tag{1.4}$$

Minimization of Eq. (1.3) proceeds using alternating minimization with the goal of finding minimizing values for $\{r_{nk}\}$ and $\{\boldsymbol{\mu}_k\}$. First, initial values for $\boldsymbol{\mu}_k$ are chosen from a random distribution and fixed, and Eq. (1.3) is minimized with respect to r_{nk}. This first step is accomplished

by assigning data points $\boldsymbol{x}_n$ to the nearest centroid. Next, r_{nk} is fixed and Eq. (1.3) minimized by setting the derivative of J with respect to $\boldsymbol{\mu}_k$ to zero:

$$2\sum_{n=1}^{N} r_{nk}(\boldsymbol{x}_n - \boldsymbol{\mu}_k) = 0, \tag{1.5}$$

whose solution is:

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} r_{nk}\boldsymbol{x}_n}{\sum_{n=1}^{N} r_{nk}}. \tag{1.6}$$

Alternating minimization is repeated until a maximum number of iterations is reached, or until changes in cluster assignments cease.

Though the k-means clustering algorithm is guaranteed to converge, it may do so at a local optimum as Eq. (1.3) is non-convex [14, 15]. Various approaches have been proposed that mitigate the risk of local convergence, including random restarts and picking centroids that cover the data space more thoroughly [14, 32]. Further limitations of k-means relate to assumptions about the cluster model, which favors well-separated clusters with spherical shapes and balanced populations. Real data rarely have these qualities, in which case k-means may be ill-suited for clustering.

A more robust algorithm that accounts for overlapping and anisotropic distributions of clusters is Gaussian mixture model (GMM) clustering [14, p. 430]. GMM clustering seeks to fit K linearly superimposed multivariate Gaussian distributions to the data using an expectation-maximization (EM) algorithm. Each Gaussian model has its own centroid $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{1.7}$$

where π_k are mixing coefficients that satisfy $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$. Similar to alternating minimization, the EM algorithm iteratively updates the Gaussian mixtures by estimating the likelihood that a sample belongs to each of the clusters and then updating centroid locations. Of note, k-means is a special case of GMM clustering, under the assumptions that $\boldsymbol{\Sigma}_k = \mathbf{I}$ and

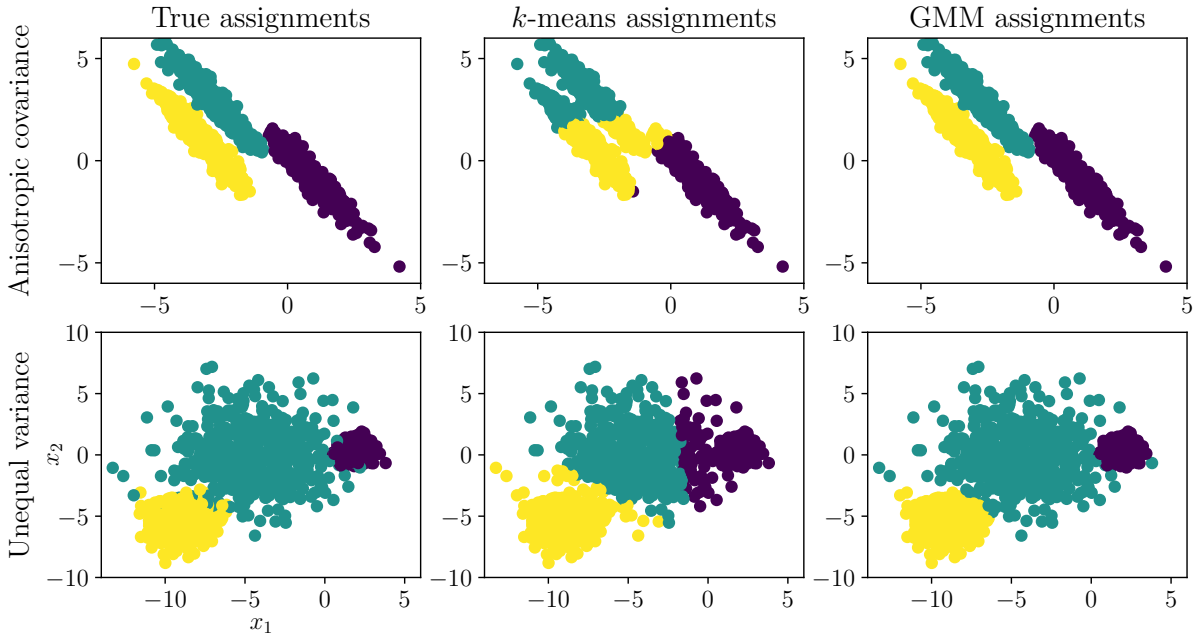

Figure 1.1. Two-dimensional data with three distinct clusters with (top) anisotropic covariances and (bottom) isotropic covariances with differing length scales. (left) True, (middle) k-means, and (right) Gaussian mixture model label assignments.

$\pi_k = 1/K$ [15, p. 728].

Figure 1.1 illustrates k-means and GMM clustering for two data sets in an example adopted from [33]. Each data set contains two dimensions and three clusters. In the top row, data with anisotropic covariance are shown. Due to the anisotropic covariance, k-means fails to properly assign labels to the data. The GMM is a better cluster model for this data, as the covariance of the clusters is one of the estimated parameters. In the bottom row of Fig. 1.1, data with isotropic but differing covariance length scales are shown. Here, too, k-means is unable to correctly label the smallest cluster, whereas GMM succeeds in estimating each of the clusters.

The most important and often most challenging parameter to choose for k-means and GMM clustering is the number of clusters K. The consequences of picking the wrong value for K are illustrated in Fig. 1.2, where data with three distinct groupings are forced into two clusters with k-means. Statistical approaches to selecting the optimal value for K include the gap statistic [34], silhouette coefficient [35], and Bayesian information criteria [15, p. 724]. While these

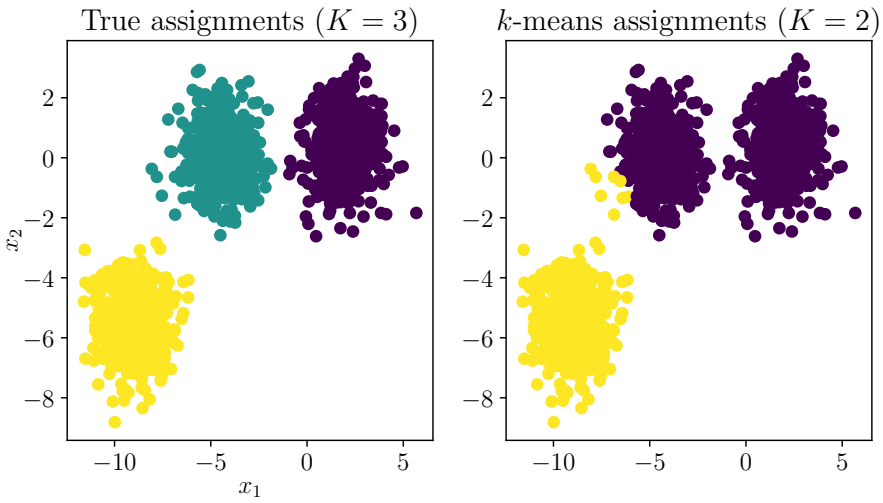

Figure 1.2. Two-dimensional data in three distinct clusters with (left) true label assignments and (right) k-means label assignments with an incorrect number of clusters chosen.

techniques can be useful tools for evaluating potential values for K, Chapter 2 [16] demonstrates how careful empirical analysis of clustering results is equally important to selecting the optimal value for K.

1.1.2 Dimensionality reduction with autoencoders

Seismic and acoustic 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). Directly clustering these high-dimensional data is vulnerable to the "curse of dimensionality" [36, 14, 15, 37], 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, making clustering in high dimensions challenging and unreliable [38, 39]. Figure 1.3 is adapted from [15, sec. 16.1.2] and illustrates the relationship between dimensionality and distance for a hypercube. As dimensionality increases, to reach an equivalent volume of the hypercube, greater distances are required. As a result, most data appear to reside near the edge of the hypercube, making distinctions in distance between points more challenging.

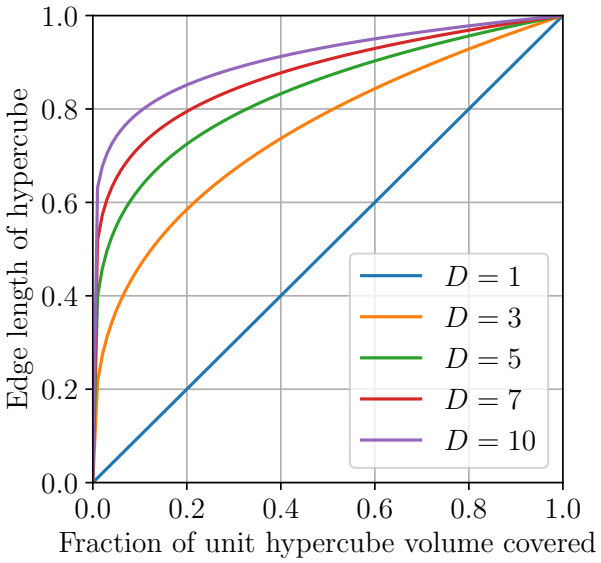

Figure 1.3. Hypercube edge lengths required to cover the fraction of the unit hypercube volume for various dimensions D.

In Chapter 2 [16], a supervised machine learning model known as an *autoencoder* reduces the dimensionality of seismic spectrograms by learning and embedding the salient information contained within the data into a latent feature space. The autoencoder model is a type of neural network and consists of three components: an *encoder*, a *bottleneck*, and a *decoder* [40, 15, 41]. The encoder f_θ provides a nonlinear mapping of data from a data space X to the bottleneck, where the latent space is contained, by $f_\theta : X \to Z$; θ are autoencoder parameters that are learned through training. The decoder reverses the encoder operation by attempting to reconstruct X from Z by $g_\theta : Z \to X'$, where X' is the reconstructed version of X. The error between X and X' is then used to update the parameters θ, and the process repeats until the autoencoder is trained. The overall mapping of the autoencoder is summarized as:

$$F_\theta : X \to Z \to X', \quad F_\theta = g_\theta \circ f_\theta. \tag{1.8}$$

For this dissertation, the nonlinear mappings of Eq. (1.8) are implemented as a deep neural network consisting of convolutional layers in the encoder and convolutional transpose layers in

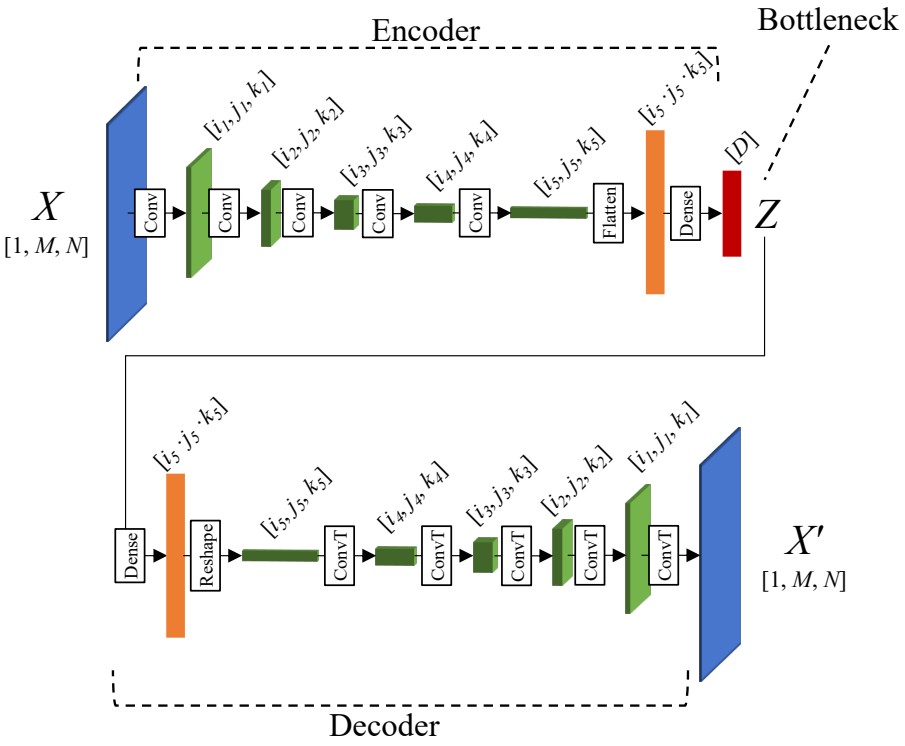

Figure 1.4. Convolutional autoencoder architecture; dimensions of each layer are shown in brackets.

the decoder. An example of a convolutional autoencoder architecture is given in Fig. 1.4.

Once trained, the embeddings contained within the bottleneck Z represent the salient features of the data set. Clustering algorithms are then applied to the lower-dimensional latent space Z, where they perform more effectively than if they were applied to the original data space X.

1.1.3 Gaussian processes

Gaussian process (GP) regression, also known as kriging, is a computationally tractable method for quantifying uncertainty and has been extensively utilized in geophysical applications [42] and recently for sound field reconstruction and prediction [43, 44, 45, 46]. The following derivations follow [47] and [15, ch. 17]. Given a distribution over functions which have the form $f : \mathscr{X} \to \mathbb{R}$, where $\mathscr{X}^D$ is a D-dimensional parameter domain, a Gaussian process (GP) is

defined as a collection of jointly Gaussian function values $\mathbf{f}$ evaluated at a set of $M > 0$ inputs $[\mathbf{x}_1, \ldots, \mathbf{x}_M]$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}_{ij}$:

$$\mathbf{f} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_M)] = GP(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{ij}) \tag{1.9}$$

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{f}] = [\mu(\mathbf{x}_1), \ldots, \mu(\mathbf{x}_M)] \tag{1.10}$$

$$\boldsymbol{\Sigma}_{ij} = \mathscr{K}(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \mathbf{I}. \tag{1.11}$$

By definition, the GP holds for unseen data, such as when M includes N training points and N^* unseen test points, i.e., $M = N + N^*$. Thus, fitting a GP is a type of regression; moreover, a fitted GP model provides both expected values and uncertainty quantification at unseen test points [15, sec. 17.2], [47, ch. 2]. For simplicity and illustration, consider a set of noise-free observations $\mathscr{D} = \{(\mathbf{x}_n, y_n) : 1 : N\} = \{\mathbf{X}, \mathbf{f}_X\}$, where $y_n = f(\mathbf{x}_n)$. A GP either returns $f(\mathbf{x})$ with no uncertainty if the point is within $\mathscr{D}$ or interpolates data not contained in $\mathscr{D}$. Expected values and uncertainty at a test set $\mathbf{X}_* = \{\mathbf{x}_n : n = 1 : N_*\}$ of unseen points are obtained from the joint distribution of the GP:

$$p(\mathbf{f}, \mathbf{f}_* | \mathbf{X}, \mathbf{X}_*) = \begin{bmatrix} \mathbf{f}_X \\ \mathbf{f}_* \end{bmatrix} = \mathscr{N}\left(\begin{bmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_* \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{X,X} & \mathbf{K}_{X,*} \\ \mathbf{K}_{X,*}^\mathsf{T} & \mathbf{K}_{*,*} \end{bmatrix} \right), \tag{1.12}$$

where $\boldsymbol{\mu}_X$ and $\boldsymbol{\mu}_*$ are the mean functions at $\mathbf{X}$ and $\mathbf{X}_*$; and

$$\mathbf{K}_{X,X} = \mathscr{K}(\mathbf{X}, \mathbf{X})_{N \times N} \tag{1.13}$$

$$\mathbf{K}_{X,*} = \mathscr{K}(\mathbf{X}, \mathbf{X}_*)_{N \times N_*} \tag{1.14}$$

$$\mathbf{K}_{*,*} = \mathscr{K}(\mathbf{X}_*, \mathbf{X}_*)_{N_* \times N_*}, \tag{1.15}$$

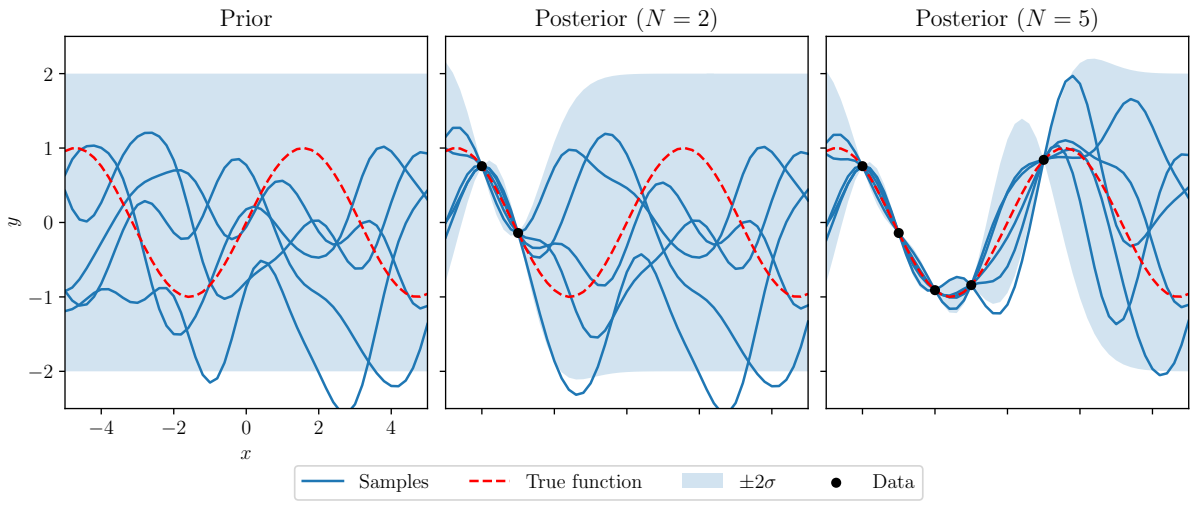

Figure 1.5. Functions (blue lines) sampled from a Gaussian process with an RBF kernel; uncertainty (blue shaded) and the true underlying function (red dotted line) are also shown. (left) A zero-mean, unity-variance prior; (middle) function samples from a GP fit with two function evaluations (black dots) and (right) five function evaluations.

where $\mathscr{K}$ is a kernel function that measures the similarity between two points. The posterior of the GP is:

$$p(\mathbf{f}_*|\mathscr{D},\mathbf{X}_*) = \mathscr{N}(\mathbf{f}_*|\boldsymbol{\mu}_{*|X},\boldsymbol{\Sigma}_{*|X}) \tag{1.16}$$

$$\boldsymbol{\mu}_{*|X} = \boldsymbol{\mu}_* + \mathbf{K}_{X,*}^{\mathsf{T}}\hat{\mathbf{K}}_{X,X}^{-1}(\mathbf{y}-\boldsymbol{\mu}_X) \tag{1.17}$$

$$\boldsymbol{\Sigma}_{*|X} = \mathbf{K}_{*,*} - \mathbf{K}_{X,*}^{T}\hat{\mathbf{K}}_{X,X}^{-1}\mathbf{K}_{X,*}. \tag{1.18}$$

Figure 1.5 illustrates several functions sampled from $p(f)$ using a radial basis function (RBF) kernel and zero mean. The true underlying function is shown in red. As the true function is evaluated and observations are added to $\mathscr{D}$, the set of possible functions drawn from the GP posterior, $p(f|\mathscr{D})$, are increasingly constrained. Points that have been evaluated have no uncertainty, whereas regions farthest from observed data have the greatest uncertainty. This is illustrated in Fig. 1.6, which shows slices of the GP at three different points in $\mathscr{X}$.

Equations (1.13)-(1.15) refer to a kernel function $\mathscr{K}$ that measures the similarity between two points in $\mathscr{X}$. The selection of a kernel function is an important choice when using GPs,

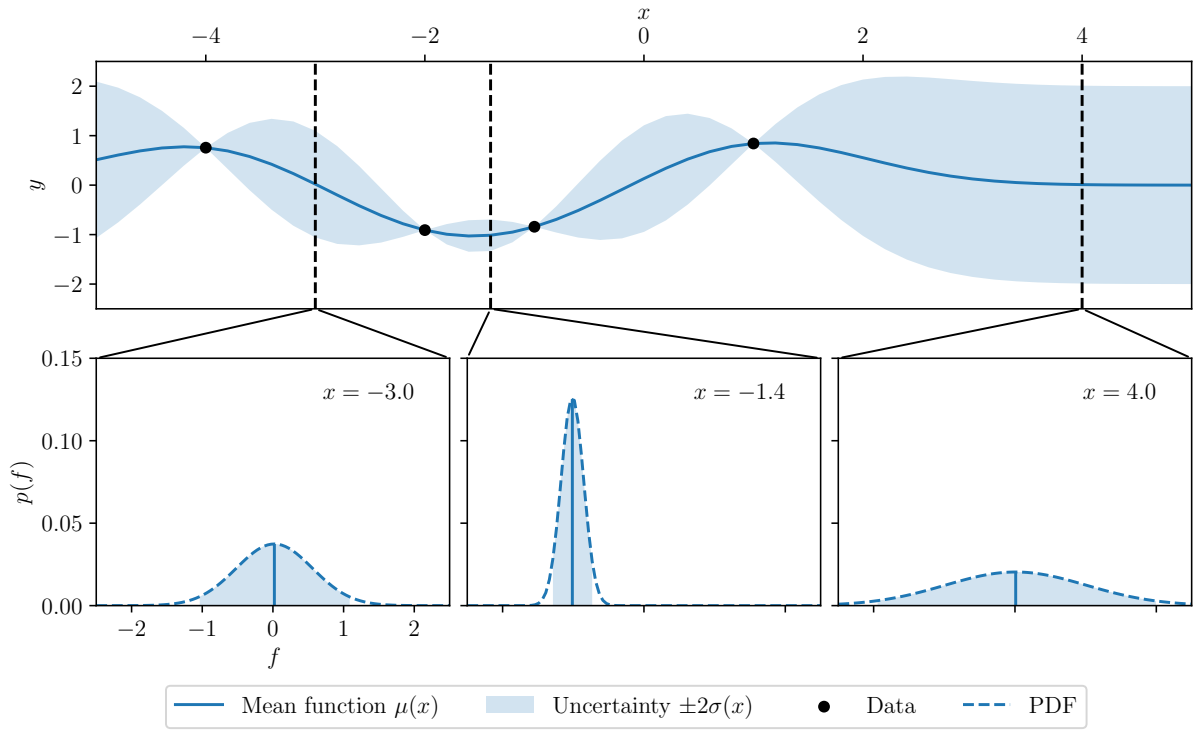

Figure 1.6. (top) A Gaussian process fit with an RBF kernel; (bottom) Slices of the GP mean and covariance functions at three points.

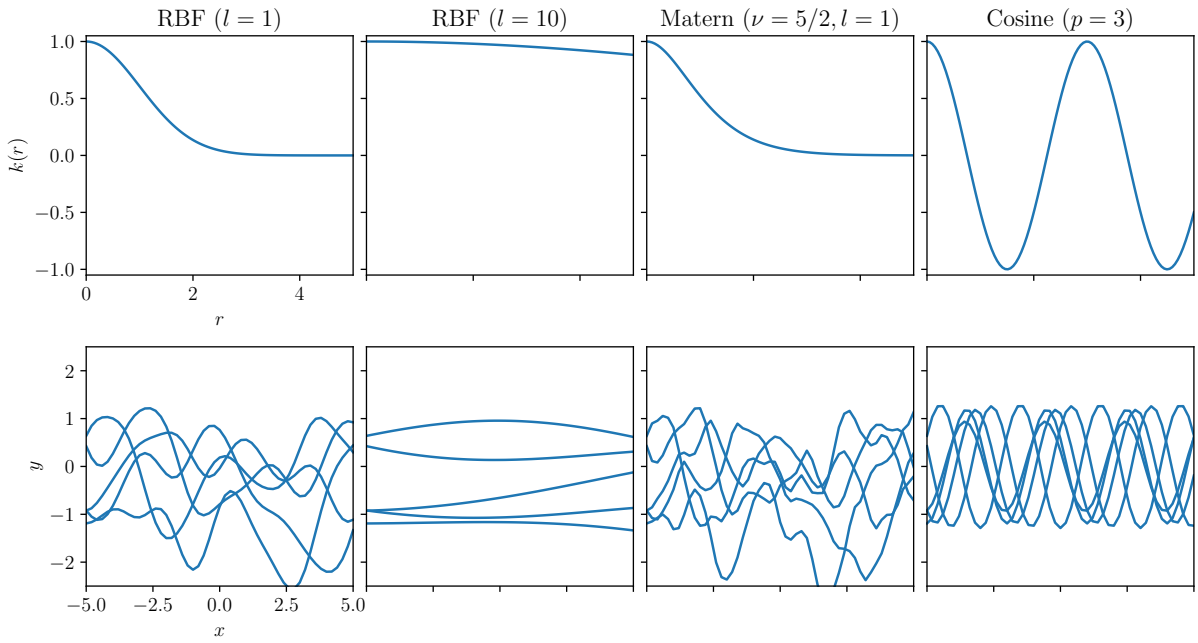

Figure 1.7. (top) Kernel functions and (bottom) associated functions drawn from $p(f)$.

as it controls the shape of the mean and covariance functions. Moreover, kernel functions contain hyperparameters $\boldsymbol{\theta}$ that must be optimized to properly reflect the observed data and provide suitable predictions in unobserved regions of $\mathscr{X}$. Figure 1.7 illustrates four kernel function implementations and associated functions drawn from $p(f)$. Using the one-dimensional illustrations as an example and defining the distance between two points in $\mathscr{X}$ as $r = \|\mathbf{x} - \mathbf{x}'\|$, the depicted kernel functions are [15, sec. 17.1]:

1. Radial basis function (RBF) kernel:

$$\mathscr{K}(r; l, \sigma_y^2) = \sigma_y^2 \exp\left(\frac{-r^2}{2l^2}\right), \tag{1.19}$$

   where σ_y^2 is the observational noise variance and l is the length scale.

2. Matern kernel:

$$\mathscr{K}(r; \nu, l, \sigma_y^2) = \sigma_y^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} r}{l}\right)^{\nu} J_\nu \left(\frac{\sqrt{2\nu} r}{l}\right), \tag{1.20}$$

   where ν is a roughness parameter and J is the modified Bessel function.

3. Cosine kernel:

$$\mathscr{K}(r; p, \sigma_y^2) = \sigma_y^2 \exp\left(2\pi \frac{r}{p}\right), \tag{1.21}$$

   where p is the period.

Kernel hyperparameters like length scale l control how rapidly functions vary with changes in r. For example, in Fig. 1.7, RBF kernels with $l = 1$ and $l = 10$ are shown; functions drawn from the former vary more rapidly than those from the latter. Observational noise variance, i.e., allowing for noise in function evaluations such that $y_n = f(\mathbf{x}_n) + \varepsilon$, $\varepsilon \sim \mathscr{N}(0, \sigma_y^2)$, determines how much uncertainty is present, including at points that belong to $\mathscr{D}$. The choice of kernel function can also be considered a hyperparameter [47, ch. 5]: the Matern kernel permits functions with both slowly and rapidly varying features, and the cosine kernel results in periodic functions.

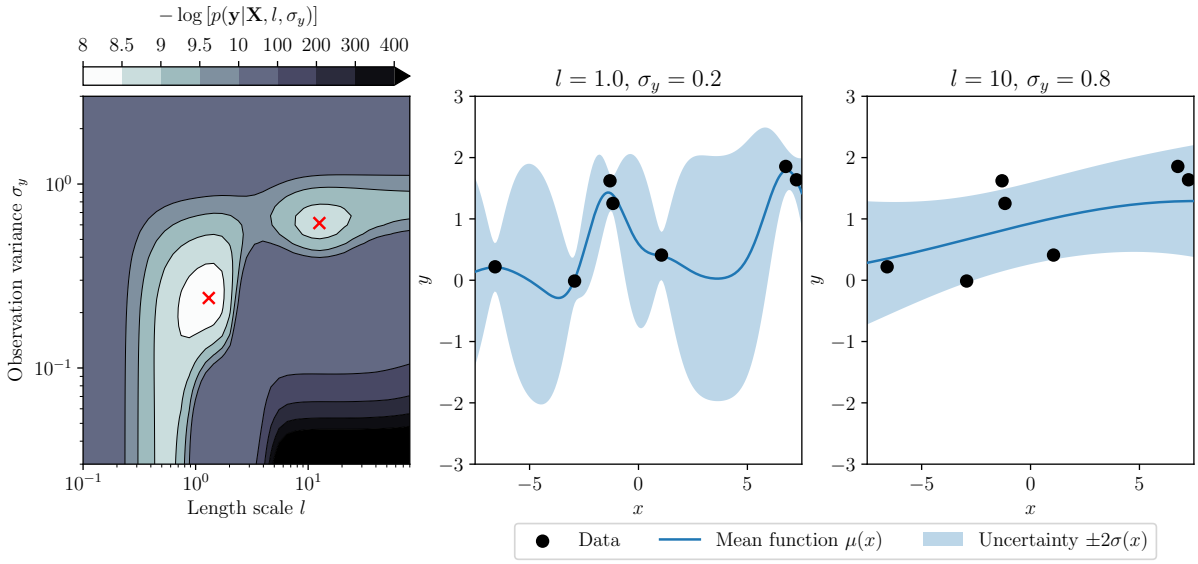

Figure 1.8. (left) Negative marginal log likelihood ($\log p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})$) for RBF kernel hyperparameters; GP regressions resulting from (middle) optimal and (right) suboptimal hyperparameter estimates.

Estimation of the kernel function hyperparameters is the critical step in GP regression. Gradient-based methods offer fast estimation by adopting an empirical Bayesian approach, which maximizes the marginal likelihood of the observations [47, ch. 5], [15, sec. 17.2.6]:

$$p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f},\mathbf{X})p(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})d\mathbf{f}. \tag{1.22}$$

In this dissertation, two gradient-based optimization methods are implemented. Chapters 4 and 5 use L-BFGS-B, a quasi-Newtonian algorithm which emulates gradient descent with momentum [48, 49], and Chapter 6 uses AdamW, an algorithm related to stochastic gradient descent [50, 51]. Figure 1.8 is adopted from [15, Fig. 17.9] and illustrates hyperparameter estimation using gradient descent optimization on the negative marginal log-likelihood surface for an RBF kernel with $\boldsymbol{\theta} = [l, \sigma_n]$. Two local minima are present, and depending on the initialization of the gradient descent algorithm, the optimization can result in an optimal fit (middle panel) or suboptimal (right panel).

1.1.4 Bayesian optimization

Bayesian optimization (BO) is a global optimization strategy that seeks to maximize an expensive-to-evaluate function about which little or nothing may be known [23, 24, 25, 26]. The objective is to find the D-dimensional parameters $\hat{\mathbf{x}} \in \mathscr{X}^D$ that maximize $f : \mathscr{X} \to \mathbb{R}$:

$$\hat{\mathbf{x}} = \underset{\mathbf{x} \in \mathscr{X}^D}{\arg\max}\, f(\mathbf{x}). \tag{1.23}$$

BO consists of the following steps:

1. Generate a data set of observations $\mathscr{D} = \{(\mathbf{x}_n, y_n) : n = 1 : N\}$ by evaluating $y = f(\mathbf{x})$ with N points drawn from $\mathscr{X}$.
2. Fit a GP surrogate model to the observed data $\mathscr{D}$ (Sec. 1.1.3).
3. Optimize an acquisition function to suggest and evaluate a new point $\mathbf{x}$; append the observation to $\mathscr{D}$.
4. Repeat steps 2 and 3 until a fixed budget of function evaluations has been expended.

To minimize the number of evaluations of f required to satisfy Eq. 1.23, a heuristic function called an *acquisition function* probabilistically guides the search by taking the GP surrogate model as its input and returning a candidate point $\mathbf{x}$ that will be evaluated upon the next iteration of BO. The candidate point is determined through optimization of the acquisition function, which is typically defined to balance exploration of the parameter space (sampling in regions with high uncertainty) with exploitation (sampling in regions of good performance) [23, 24]. Consider, for example, the upper confidence bound acquisition function [24, 52]:

$$\alpha_{\text{UCB}}(f(\mathbf{x})) = \mu(\mathbf{x}) + \kappa\sigma(\mathbf{x}), \tag{1.24}$$

where μ and σ are the mean and covariance functions of the GP surrogate model, respectively, and κ is a hyperparameter that controls the contributions of each term. Regions of high performance

(large values of the objective function f) compete against regions of high uncertainty; in this way, the algorithm is attempting to prioritize between exploiting regions of high performance and exploring regions with few observations and high uncertainty. The degree to which UCB favors exploitation vs. exploration is controlled by κ. The study and development of acquisition functions, as well as their behavior and convergence proofs, remain active areas of research.

Due to its ability to incorporate observations of the objective function f in its decision-making about where to sample next, BO can efficiently find regions of optimal performance, providing an advantage over exhaustive and random search. However, like all optimization algorithms, BO is susceptible to converging on local optima, though various techniques have been implemented to make BO more robust in the optimization of multi-modal objective functions.

1.2 Dissertation overview

Chapter 2 [16] investigates the application of deep clustering for exploratory data analysis of continuous seismic data collected on the Ross Ice Shelf, Antarctica from 2014-2016 [53]. The large, two-dimensional array spanning the ice shelf consisted of 34 seismic stations recording at high sampling rates (100 and 200 Hz). A detection algorithm [54] detected more than 530,000 seismic events, whose spectrograms were reduced to a low-dimensional latent space using a convolutional autoencoder. Next, clustering was performed on the data in the latent space using GMM clustering [40, 15] and DEC [18]. GMM yielded satisfactory results with high computational efficiency, while DEC was computationally expensive to implement and difficult to interpret due to distortions in the latent space. Clustering results for the two years of data were analyzed and diurnal, seasonal, and interannual patterns were identified and correlated with potential source mechanisms such as ocean wave impacts at the shelf terminus, tidal activity near grounding lines, and high melting rates due to El Niño.

Chapter 3 [55] investigates acoustic data collected during an oceanographic training cruise conducted in sea ice north of Svalbard, Norway in June 2021. Acoustic data collected

during the cruise revealed a high amount of biological activity, including various fish and marine mammals. A hydrophone at an ice station recorded the detonation of expired explosives, providing an opportunity to measure the depth of the ocean using the impulsive signal. Multiple arrivals and the modal structure of the sound are visible in the spectrogram of the recording.

Chapter 4 [27] investigates acoustic source localization using Bayesian optimization with Gaussian processes [26]. Under this construct, the objective is to estimate source location in range and depth as accurately and with as few objective function evaluations as possible. Using the matched field processing (MFP) localization framework [4], the objective function is the Bartlett power ambiguity surface, which is modeled as a GP surrogate model [47]. Bayesian optimization is performed with two different acquisition functions—expected improvement (EI) [24] and quasi-Monte Carlo EI [42, 56, 57, 58]—which are evaluated and compared against conventional grid search, quasi-random search [59], and sparse Bayesian learning (SBL) [60, 61, 62, 63].

Chapter 5 [28] extends the method demonstrated in Chapter 4 by adding receiver array tilt to the parameter search space. Vertical line arrays (VLA) often tilt due to currents, which leads to model mismatch in MFP. Various studies have confirmed that array tilt is a sensitive parameter in acoustic source localization and geoacoustic inversion [64, 65, 62, 63, 66, 67, 11, 68]. By allowing BO to jointly estimate source localization and array tilt, better correlations between predicted and observed data are obtained, leading to more accurate estimates of source localization and array tilt.

Chapter 6 demonstrates Bayesian optimization in a higher-dimensional setting, jointly estimating source localization and geoacoustic parameters. Similar performance characteristics observed in chapters 4 and 5 are observed for geoacoustic inversion, with the quality of parameter estimates consistent with forward model sensitivity.

1.3 References

[1] K. Aki and P. G. Richards, *Quantitative Seismology*. University Science Books, 2002.

[2] A. T. Ringler, R. E. Anthony, R. C. Aster, C. J. Ammon, S. Arrowsmith, H. Benz, C. Ebeling, A. Frassetto, W.-Y. Kim, P. Koelemeijer, H. C. P. Lau, V. Lekić, J. P. Montagner, P. G. Richards, D. P. Schaff, M. Vallée, and W. Yeck, "Achievements and Prospects of Global Broadband Seismographic Networks After 30 Years of Continuous Geophysical Observations," *Reviews of Geophysics*, vol. 60, p. e2021RG000749, Sept. 2022.

[3] W. H. Munk and P. F. Worcester, "Ocean Acoustic Tomography," *Oceanography*, vol. 1, pp. 8–10, July 1988.

[4] A. B. Baggeroer, W. A. Kuperman, and P. N. Mikhalevsky, "An overview of matched field methods in ocean acoustics," *IEEE J. Ocean. Eng.*, vol. 18, pp. 401–424, Oct. 1993.

[5] W. Munk, P. Worcester, and C. Wunsch, *Ocean Acoustic Tomography*. Cambridge Monographs on Mechanics, Cambridge University Press, 2009.

[6] N. R. Chapman and E. C. Shang, "Review of Geoacoustic Inversion in Underwater Acoustics," *J. Theor. Comp. Acout.*, vol. 29, p. 2130004, Sept. 2021.

[7] E. A. Podolskiy and F. Walter, "Cryoseismology," *Reviews of Geophysics*, vol. 54, pp. 708–758, Dec. 2016.

[8] R. C. Aster and J. P. Winberry, "Glacial seismology," *Rep. Prog. Phys.*, vol. 80, p. 126801, Dec. 2017.

[9] M. J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M. A. Roch, S. Gannot, and C.-A. Deledalle, "Machine learning in acoustics: Theory and applications," *The Journal of the Acoustical Society of America*, vol. 146, pp. 3590–3628, Nov. 2019.

[10] Q. Kong, D. T. Trugman, Z. E. Ross, M. J. Bianco, B. J. Meade, and P. Gerstoft, "Machine Learning in Seismology: Turning Data into Insights," *Seismological Research Letters*, vol. 90, pp. 3–14, Jan. 2019.

[11] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *J. Acoust. Soc. Am.*, vol. 152, pp. 107–151, July 2022.

[12] S. M. Mousavi and G. C. Beroza, "Deep-learning seismology," *Science*, vol. 377, Aug. 2022.

[13] S. M. Mousavi and G. C. Beroza, "Machine Learning in Earthquake Seismology," *Annu. Rev. Earth Planet. Sci.*, vol. 51, pp. 105–129, May 2023.

[14] C. Bishop, *Pattern Recognition and Machine Learning*. Information Science and Statistics, Springer-Verlag New York, 1 ed., 2006.

[15] K. P. Murphy, *Probabilistic Machine Learning: An Introduction*. Cambridge, MA: MIT Press, 2022.

[16] W. F. Jenkins II, P. Gerstoft, M. J. Bianco, and P. D. Bromirski, "Unsupervised Deep Clustering of Seismic Data: Monitoring the Ross Ice Shelf, Antarctica," *JGR Solid Earth*, vol. 126, Aug. 2021.

[17] E. Ozanich, A. Thode, P. Gerstoft, L. A. Freeman, and S. Freeman, "Deep embedded clustering of coral reef bioacoustics," *J. Acoust. Soc. Am.*, pp. 2587–2601, 2021.

[18] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised Deep Embedding for Clustering Analysis," *Proceedings of the 33rd international conference on machine learning*, p. 10, 2016.

[19] S. M. Mousavi, W. Zhu, W. Ellsworth, and G. Beroza, "Unsupervised Clustering of Seismic Signals Using Deep Convolutional Autoencoders," *IEEE Geosci. Remote Sensing Lett.*, vol. 16, pp. 1693–1697, Nov. 2019.

[20] L. Seydoux, R. Balestriero, P. Poli, M. de Hoop, M. Campillo, and R. Baraniuk, "Clustering earthquake signals and background noises in continuous seismic data with unsupervised deep learning," *Nat Commun*, vol. 11, p. 3972, Dec. 2020.

[21] D. Snover, C. W. Johnson, M. J. Bianco, and P. Gerstoft, "Deep Clustering to Identify Sources of Urban Seismic Noise in Long Beach, California," *Seismological Research Letters*, vol. 92, pp. 1011–1022, Mar. 2021.

[22] R. C. Aster, B. Borchers, and C. H. Thurber, *Parameter Estimation and Inverse Problems*. Cambridge, MA: Elsevier, 2018.

[23] D. R. Jones, M. Schonlau, and W. J. Welch, "Efficient Global Optimization of Expensive Black-Box Functions," *Journal of Global Optimization*, vol. 13, no. 4, pp. 455–492, 1998.

[24] D. R. Jones, "A Taxonomy of Global Optimization Methods Based on Response Surfaces," *J. Global Optim.*, vol. 21, no. 4, pp. 345–383, 2001.

[25] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, "Taking the Human Out of the Loop: A Review of Bayesian Optimization," *Proc. IEEE*, vol. 104, pp. 148–175, Jan. 2016.

[26] S. Greenhill, S. Rana, S. Gupta, P. Vellanki, and S. Venkatesh, "Bayesian Optimization for Adaptive Experimental Design: A Review," *IEEE Access*, vol. 8, pp. 13937–13948, 2020.

[27] W. F. Jenkins II, P. Gerstoft, and Y. Park, "Bayesian optimization with Gaussian process surrogate model for source localization," *J Acoust. Soc. Am.*, vol. 154, pp. 1459–1470, Sept. 2023.

[28] W. F. Jenkins II and P. Gerstoft, "Bayesian optimization with Gaussian processes for robust localization," *Submitted to IEEE Int. Conf. Acoust., Speech, Signal Process.*, Sept. 2023.

[29] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, (Berkeley, Calif.), pp. 281–297, University of California Press, 1967.

[30] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," *Applied Statistics*, vol. 28, no. 1, p. 100, 1979.

[31] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inform. Theory*, vol. 28, pp. 129–137, Mar. 1982.

[32] D. Arthur and S. Vassilvitskii, "K-Means++: The advantages of careful seeding," in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, (USA), pp. 1027–1035, Society for Industrial and Applied Mathematics, 2007.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[34] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *J Royal Statistical Soc B*, vol. 63, pp. 411–423, May 2001.

[35] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, Nov. 1987.

[36] R. E. Bellman, *Adaptive Control Processes: A Guided Tour*. Rand Corporation, 1961.

[37] C. C. Aggarwal and C. K. Reddy, eds., *Data Clustering: Algorithms and Applications*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, Boca Raton: Chapman and Hall/CRC, 2014.

[38] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the Surprising Behavior of Distance Metrics in High Dimensional Space," in *Database Theory — ICDT 2001* (G. Goos, J. Hartmanis, J. van Leeuwen, J. Van den Bussche, and V. Vianu, eds.), vol. 1973, pp. 420–434, Berlin, Heidelberg: Springer Berlin Heidelberg, 2001.

[39] M. Steinbach, L. Ertöz, and V. Kumar, "The Challenges of Clustering High Dimensional Data," in *New Directions in Statistical Physics: Econophysics, Bioinformatics, and Pattern Recognition* (L. T. Wille, ed.), pp. 273–309, Berlin, Heidelberg: Springer Berlin Heidelberg, 2004.

[40] G. E. Hinton, "Reducing the Dimensionality of Data with Neural Networks," *Science*, vol. 313, pp. 504–507, July 2006.

[41] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016.

[42] D. Ginsbourger, R. Le Riche, and L. Carraro, "Kriging Is Well-Suited to Parallelize Optimization," in *Computational Intelligence in Expensive Optimization Problems*, vol. 2, pp. 131–162, Berlin, Heidelberg: Springer, 2010.

[43] D. Caviedes-Nozal, N. A. B. Riis, F. M. Heuchel, J. Brunskog, P. Gerstoft, and E. Fernandez-Grande, "Gaussian processes for sound field reconstruction," *J. Acoust. Soc. Am.*, vol. 149, pp. 1107–1119, Feb. 2021.

[44] Z.-H. Michalopoulou, P. Gerstoft, and D. Caviedes-Nozal, "Matched field source localization with Gaussian processes," *JASA Express Lett.*, vol. 1, p. 064801, June 2021.

[45] Z.-H. Michalopoulou and P. Gerstoft, "Inversion in an uncertain ocean using Gaussian processes," *J. Acoust. Soc. Am.*, vol. 153, pp. 1600–1611, Mar. 2023.

[46] I. D. Khurjekar, P. Gerstoft, C. F. Mecklenbräuker, and Z.-H. Michalopoulou, "Direction-of-Arrival Estimation Using Gaussian Process Interpolation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 1–5, 2023.

[47] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press, 2006.

[48] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimization," *ACM Trans. Math. Software*, vol. 23, pp. 550–560, Dec. 1997.

[49] J. Nocedal and S. J. Wright, *Numerical Optimization*. Springer Series in Operations Research, New York: Springer, 2nd ed., 2006.

[50] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980 [cs]*, Jan. 2017.

[51] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *Int. Conf. Learning Representations*, 2019.

[52] N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger, "Information-Theoretic Regret Bounds for Gaussian Process Optimization in the Bandit Setting," *IEEE Trans. Inf. Theory*, vol. 58, no. 5, pp. 3250–3265, 2012.

[53] P. D. Bromirski, A. Diez, P. Gerstoft, R. A. Stephen, T. Bolmer, D. A. Wiens, R. C. Aster, and A. Nyblade, "Ross ice shelf vibrations," *Geophys. Res. Lett.*, vol. 42, pp. 7589–7597, Sept. 2015.

[54] R. Allen, "Automatic phase pickers: Their present use and future prospects," *Bulletin of the Seismological Society of America*, vol. 72, pp. S225–S242, Dec. 1982.

[55] W. Jenkins, H. Johnson, S. Vakhutinsky, M. N. Helmberger, E. Storheim, H. Sagen, and S. Sandven, "Analysis of underwater acoustic data collected under sea ice during the Useful Arctic Knowledge 2021 cruise," in *Proceedings of Meetings on Acoustics*, (Southampton, UK), p. 070003, Acoustical Society of America, 2022.

[56] J. T. Wilson, F. Hutter, and M. P. Deisenroth, "Maximizing acquisition functions for Bayesian optimization," in *Advances in Neural Information Processing Systems*, vol. 32, 2018.

[57] J. Wang, S. C. Clark, E. Liu, and P. I. Frazier, "Parallel Bayesian Global Optimization of Expensive Functions," *Oper. Res.*, vol. 68, pp. 1850–1865, Nov. 2020.

[58] M. Balandat, B. Karrer, D. R. Jiang, S. Daulton, B. Letham, A. G. Wilson, and E. Bakshy, "BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization," in *Advances in Neural Information Processing Systems*, vol. 34, 2020.

[59] I. M. Sobol', "On the distribution of points in a cube and the approximate evaluation of integrals," *USSR Comp. Math. and Math. Phys.*, vol. 7, pp. 86–112, Jan. 1967.

[60] M. E. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.

[61] D. Wipf and B. Rao, "Sparse Bayesian Learning for Basis Selection," *IEEE Trans. Signal Process.*, vol. 52, pp. 2153–2164, Aug. 2004.

[62] K. L. Gemba, S. Nannuru, P. Gerstoft, and W. S. Hodgkiss, "Multi-frequency sparse Bayesian learning for robust matched field processing," *J. Acoust. Soc. Am.*, vol. 141, pp. 3411–3420, May 2017.

[63] K. L. Gemba, S. Nannuru, and P. Gerstoft, "Robust Ocean Acoustic Localization With Sparse Bayesian Learning," *IEEE J. Sel. Top. Signal Process.*, vol. 13, pp. 49–60, Mar. 2019.

[64] P. Gerstoft and C. F. Mecklenbräuker, "Ocean acoustic inversion with estimation of *a posteriori* probability distributions," *J. Acoust. Soc. Am.*, vol. 104, pp. 808–819, Aug. 1998.

[65] C. Yardim, P. Gerstoft, and W. S. Hodgkiss, "Geoacoustic and source tracking using particle filtering: Experimental results," *J. Acoust. Soc. Am.*, vol. 128, pp. 75–87, July 2010.

[66] G. Byun, F. H. Akins, K. L. Gemba, H. C. Song, and W. A. Kuperman, "Multiple constraint matched field processing tolerant to array tilt mismatch," *J. Acoust. Soc. Am.*, vol. 147, pp. 1231–1238, Feb. 2020.

[67] J. Bonnel, S. E. Dosso, J. A. Goff, Y. Lin, J. H. Miller, G. Potty, P. Wilson, and D. Knobles, "Transdimensional Geoacoustic Inversion Using Prior Information on Range-Dependent Seabed Layering," *IEEE J. Oceanic Eng.*, vol. 47, pp. 594–606, July 2022.

[68] A. Weiss, A. C. Singer, and G. W. Wornell, "Towards Robust Data-Driven Underwater Acoustic Localization: A Deep CNN Solution with Performance Guarantees for Model Mismatch," in *Proc. IEEE ICASSP*, pp. 1–5, June 2023.

Chapter 2

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

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 clustering was employed to efficiently investigate these signals. Deep clustering automatically groups 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 method uses spectrograms as input and encodes their salient features into a lower-dimensional latent representation using an autoencoder, a type of deep neural network. For comparison, two clustering methods are applied to the latent data: a Gaussian mixture model (GMM) and deep embedded clustering (DEC). 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.

2.1 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 [6]. Warming oceans are enhancing basal melting of ice shelves that reduces the buttressing of grounded ice sheets [7, 8, 9, 10], leading to increased discharge of ice into the ocean and raising sea level [11, 12, 13, 14]. With West Antarctica alone containing a sea level rise potential of 5.6 m [6], 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 [15]. 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 [16, 17, 18, 19]. In this study we present an implementation of *clustering*, a form

of unsupervised machine learning used to discover classes of similar signals within a data set [20, 21, 22], 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 [1]. The array, shown in Figure 2.1, 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 [23]. Signals and ambient noise of interest on the RIS include tidally-driven stick-slip seismicity at Whillans Ice Stream [24, 25, 26]; basal micro-earthquakes and tremor [27]; tidally and thermally driven rift fractures [28]; diurnal seismicity associated with subsurface melting [29]; wind-generated resonance in the ice [30]; flexural and plate waves generated by ocean swell, infragravity waves, and tsunami [31, 32, 33]; regional and teleseismic earthquakes [34]; and icequakes generated by ocean gravity waves [35]. Ambient seismic noise, which can be used to estimate the RIS structure [36], also contains spectra from ocean gravity waves, whose dispersion can be used to identify their source distance and origin [1, 37].

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 two unsupervised clustering methodologies to the RIS array seismic 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.

2.2 Background

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

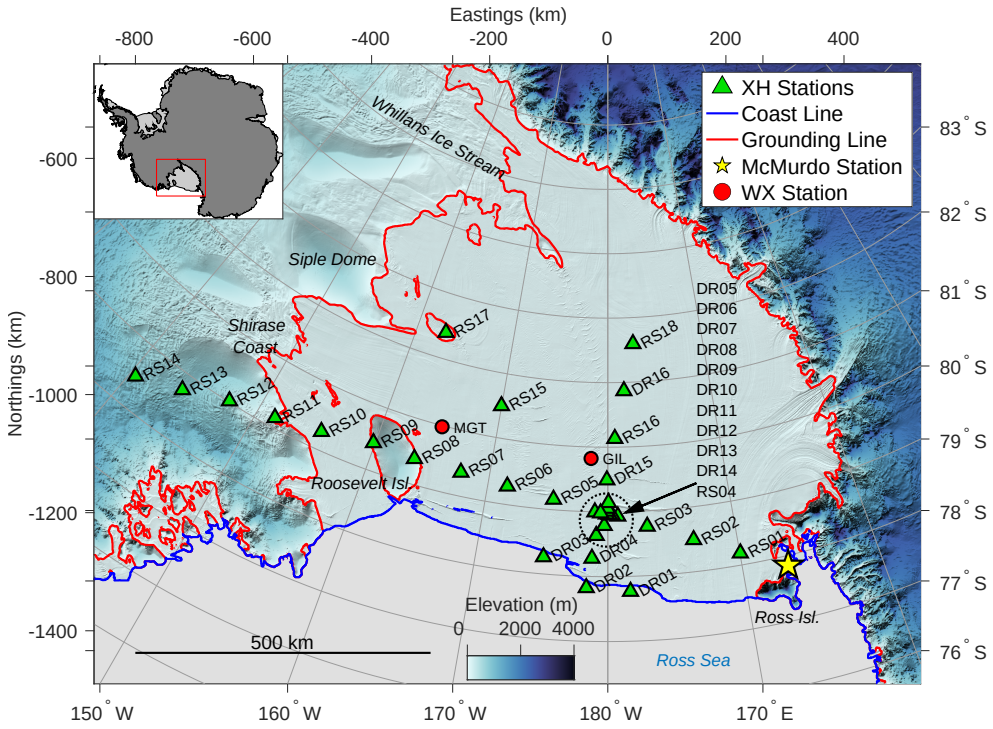

Figure 2.1. The passive broadband seismic array deployed from November 2014 to January 2017 consisted of 34 seismic stations and was deployed as part of the Ross Ice Shelf Dynamic Response to Wave-Induced Vibrations Project [1]. RIS surface elevation, ice and water layer thicknesses, and grounding and coast lines were obtained from Bedmachine [2, 3].

2.2.1 Clustering

There are numerous methods to cluster data [38], many of which have been adapted for use in seismology and geophysics [16]. A related approach based on sparse modeling, called dictionary learning, has been applied to regularizing seismic inverse problems [17, 18]. Hierarchical clustering has been used by [39] to automatically discriminate between shallow and deep earthquakes, and by [40] to more precisely localize earthquakes. Graphical clustering has been used to localize sources in a dense seismic array by [41], and by [42] to cluster seismic events in time. Distance-based clustering, like the popular k-means algorithm [43, 44], has been used by [45] to cluster seismicity based on features extracted from seismic data. [46] used k-means to define probabilistic earthquake locations as part of their convolutional neural network (CNN) detection and localization technique. [47] used Gaussian mixture model (GMM) clustering, which assumes clusters in the data exist that can be represented as linearly superimposed Gaussian distributions, enabling identification of seismic facies. [48] detected and clustered seismic signals and background noise with the use of a deep scattering neural network and GMM.

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 [49, 50, 51]. [52] and [53] presented computationally efficient techniques in which locality-sensitive hashing is used to map seismic signals into a hash table, allowing similar signals to be identified by table entry. [54] developed an approach that uses correlation-based similarity search to automatically detect and cluster repeating volcanic seismicity in continuous data. [55] adopted the method of [54] to cluster RIS array data at stations RS09, RS10, and RS11 in order to characterize tidal forcing of seismicity at these stations.

2.2.2 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 is vulnerable to the "curse of dimensionality" [56, 20, 57, 38], 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 [58, 59], dimensionality reduction remains a major focus of development [60]. It is often desirable to transform the input data to a lower-dimensional representation described by fewer, more salient features. A popular approach is to use principal component analysis (PCA), which projects higher dimensional data into lower dimensional space [61] and was used by [62] to compress seismic data to maximize feature variance.

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) [63, 57, 60]. These non-linear transformations provide greater capacity in dimension reduction, and can better model data with low-dimensional representations than, for example, PCA. 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.

2.2.3 Deep Embedded Clustering

In deep clustering, a DNN such as an autoencoder is used to reduce the dimensionality of the data. A recent deep clustering method that has shown improvement over traditional clustering techniques was developed by [64], whose *deep embedded clustering* (DEC) consists of two processes: (1) An autoencoder is trained to represent the data's salient features; and (2) the encoding layers and clustering layer are jointly optimized. [60] extended 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: [64] used a stacked denoising autoencoder [65] in their original implementation, but [66] employed autoencoders composed of CNN layers and other architectures. More recently, [67] developed an approach in which joint clustering is performed with a mixture of autoencoders, each representing a cluster, and [68] demonstrated improved performance using a clustering algorithm that is jointly optimized with the embeddings of the autoencoder.

[69] used DEC to predict whether seismic detections were local or teleseismic, and [70] demonstrated the ability of DEC to cluster anthropogenically generated seismic noise. In a similar signal processing and clustering workflow to ours, [71] compared DEC and GMM on spectrograms of acoustic data collected on a coral reef, but in their case found GMM performed better than DEC.

In this study, we implement GMM clustering in the latent feature space and compare its performance with DEC. Using RIS seismic data from December 2014 to November 2016, we identify several different classes of signals, and further demonstrate the utility of deep clustering as an exploratory tool for large, real-world seismic data sets by associating the clustering results with observed environmental factors.

2.3 Ross Ice Shelf (RIS) 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 ice shelf front 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 with frequencies of interest occurring between 3 and 20 Hz. This passband was selected to preserve impulsive signals, eliminate high-energy noise prevalent at low frequencies, and exclude resonances generated by wind at frequencies above 20 Hz. Representative types of signals detected are shown in Figure 2.2.

Seismic data from each station were processed in 24-hour segments as follows: 1) Data were linearly de-trended and tapered with a Hann window. 2) Instrument responses for all stations were removed, giving acceleration in $\mathrm{m/s^2}$. 3) Since the bandwidth of interest was from 3 to 20 Hz, data were decimated to 50 Hz, using low-pass filtering followed-by downsampling. 4) A band-pass 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. 5) A short-term average/long-term average (STA/LTA) detection algorithm [72] was used to detect impulsive signals, particularly icequakes and local earthquakes, employing an STA window of 0.5 s, LTA window of 30 s, trigger threshold of 15, and de-trigger threshold of 10. The detector was applied to data from each station from 3 December 2014 to 21 November 2016 for a total of 719 days of array data, yielding 531,407 detections.

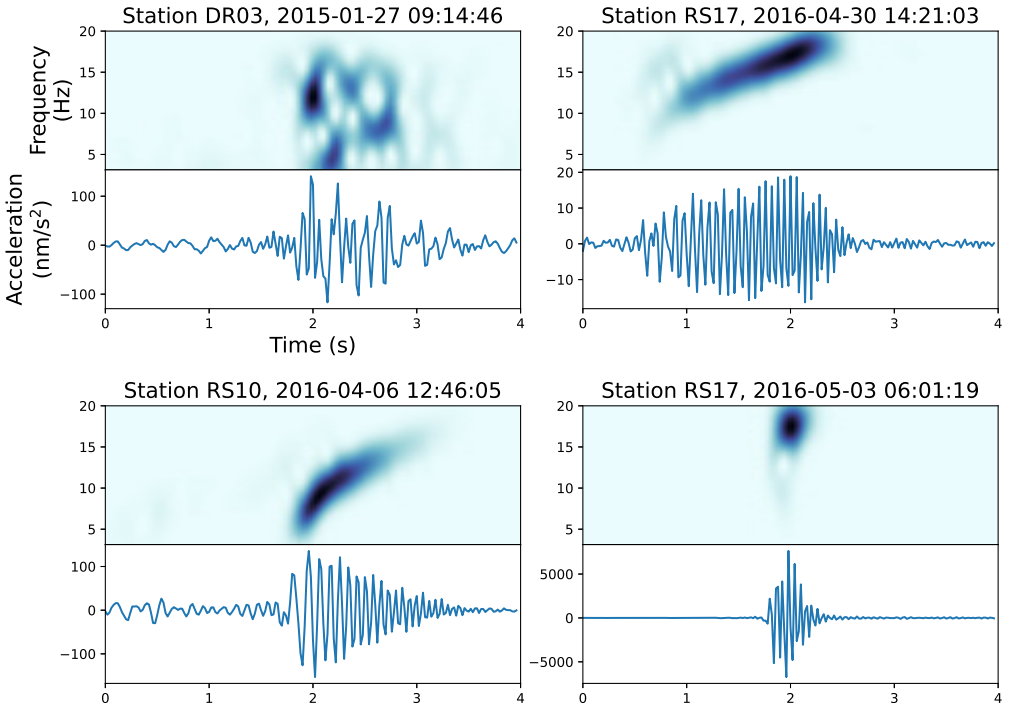

Figure 2.2. Seismic signals detected on the Ross Ice Shelf exhibited diverse characteristics with variation in time, space, and source mechanism. Shown are examples of acceleration response seismograms and their respective normalized spectrograms spanning the 3-20 Hz band that were typical for the data set. The normalized spectrograms were used as input to the deep clustering analysis.

Upon detection, a 4 s trace centered on the spectral peak of each triggered event was saved for processing. Centering the trace at the spectral peak yielded more unique clusters by preventing the clustering algorithm from labeling similar signals as different classes based only on their relation to the trigger time. 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 (samples) contained one channel of amplitude information, 87 frequency bins, and 100 time bins for a total of 8,700 features per spectrogram. To improve DNN learning, sample-wise normalization was performed by dividing each spectrogram by its vector norm [73].

2.4 Deep Clustering Implementation

The objective of deep clustering models is to first encode the input data—in this case, spectrograms of seismic signals—into a layer containing latent (lower-dimensional) features, called the *embedded* layer, and to then apply a clustering algorithm in this latent feature space. In the implementation that follows, the 8,700 features of an input 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 and transposed convolutional layers. We then describe the GMM and DEC clustering algorithms that are used in the clustering analysis.

2.4.1 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* [57]. 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.

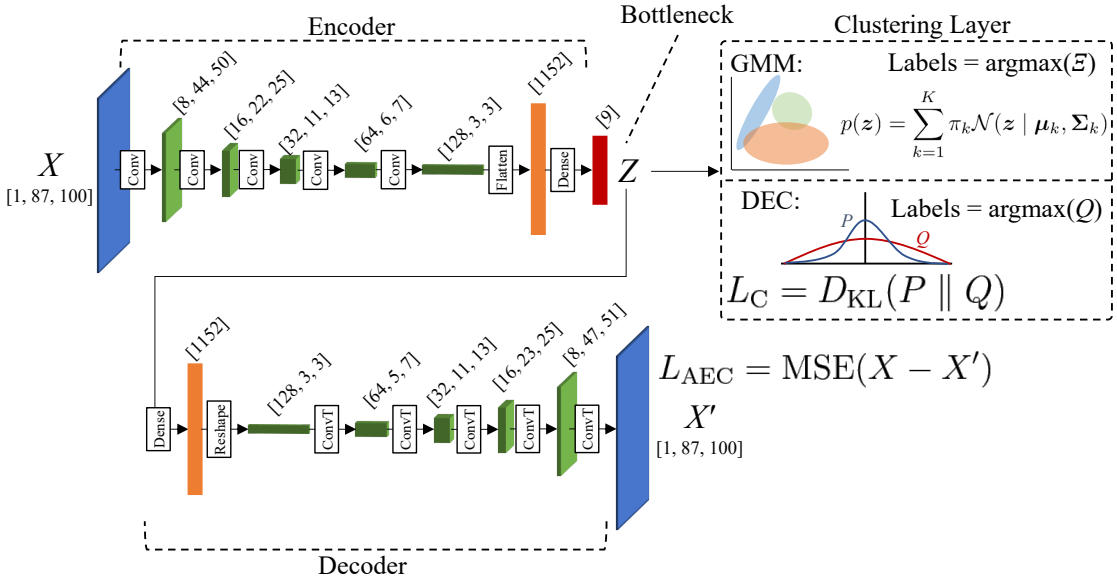

Figure 2.3. The deep clustering framework in this study uses a convolutional autoencoder that encodes the data space X into the latent feature space Z, and a decoder that recovers the original input X from Z. The mean squared error (MSE) between the input X and the reconstruction X' is used as the autoencoder loss function. The latent feature space Z lies at the bottleneck between the encoder and decoder, providing the input to the clustering layer. Gaussian mixture model (GMM) clustering labels each data sample according to its most likely cluster membership using an expectation-maximization algorithm. Deep embedded clustering (DEC) provides label assignments, and also outputs a clustering loss function that is combined with the MSE to further train the parameters that map $X \rightarrow Z \rightarrow X'$.

Consider a data set of spectrograms $\mathscr{D} = \{\boldsymbol{x}_n \in X^M\}_{n=1}^N$, where $\boldsymbol{x}_n$ is a vector representation of the n^{th} spectrogram in a data set containing N spectrograms, and the number of features in $\boldsymbol{x}_n$, 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 $\boldsymbol{x}_n$ map to their corresponding latent feature vectors by $\boldsymbol{z}_n = f_\theta(\boldsymbol{x}_n) \in Z^D$, where D is the number of embedded features, and to their reconstructions by $\boldsymbol{x}'_n = F_\theta(\boldsymbol{x}_n) \in X'$.

As the autoencoder is composed of convolutional and transposed convolutional layers, 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

$$\ell(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{M} \sum_{m=1}^{M} (x_m - x'_m)^2, \tag{2.1}$$

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

$$L_{\text{AEC}} = \frac{1}{N} \sum_{n=1}^{N} \ell(\boldsymbol{x}_n, \boldsymbol{x}'_n). \tag{2.2}$$

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, stochastic gradient descent [61] is used to update the weights. When all mini-batches have been processed, the next training epoch begins and the process is repeated. After each epoch, a subset of the data separate from the training data is used to validate the model's performance without updating the weights, yielding a validation

Table 2.1. 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]	-	-	[1152]	0
Encoded	Fully Connected	[1152]	-	ReLU	[9]	10,377
FC	Fully Connected	[9]	-	ReLU	[1152]	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**

Table 2.2. Sample Sizes and Hyperparameters used to Train the Autoencoder and Deep Embedded Clustering Model

Samples			**Hyperparameters**				
Total (N)	Training (N_{train})	Validation (N_{val})	Initial learning rate	Mini-batch size	Classes (K)	Clustering loss factor (λ)	Updates per epoch
531,407	40,000	10,000	10^{-3}	64	8	10^{-4}	10

MSE. Training is performed until a specified maximum number of epochs is reached, or stopped early if the validation MSE fails to decrease below its minimum value after ten epochs. The early stopping criterion prevents the autoencoder from overfitting the training data.

The design choice of autoencoder architecture can be informed by prior knowledge of a data set and its features, as well as practical considerations such as computational resources available. Our DNN architecture, detailed in Table 2.1, is designed to be computationally efficient, simple to construct, and robust enough to learn salient features from a noisy seismic data set. In total, θ contains 218,250 trainable parameters under this DNN architecture.

Autoencoder training is implemented using 50,000 spectrograms randomly selected

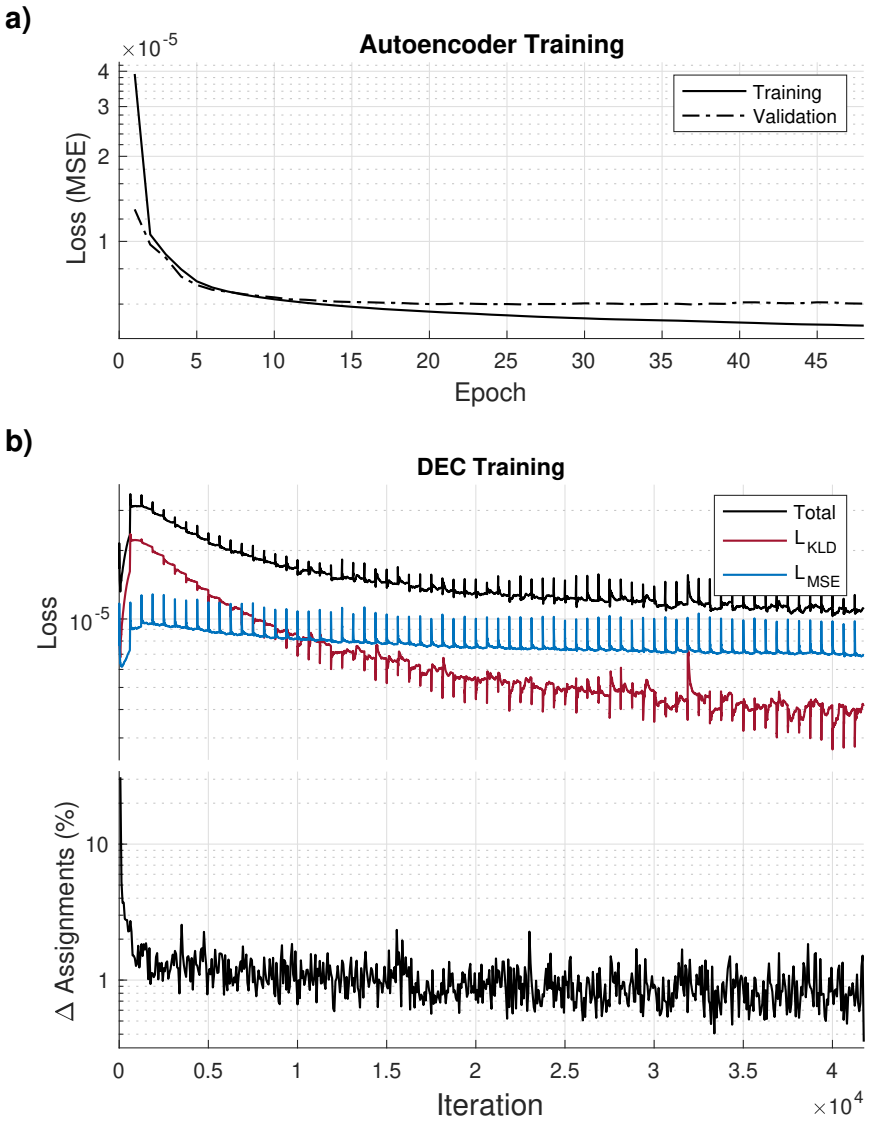

Figure 2.4. (a) Training and validation losses during autoencoder training. To avoid over-fitting the model, training is stopped when the early stopping criterion is met (in this case, at 48 epochs). (b) In the upper plot, loss curves are shown for deep embedded clustering (DEC). In the lower plot, the percentage of samples which undergo class reassignment at each update interval is shown; training is stopped once the change is less than 0.4%

without replacement from the 531,407 detections. Of the selected spectrograms, 80% are used for training and 20% for validation. The trainable parameters are optimized using the Adaptive Moment Estimation (Adam) algorithm [74]. In training, there are two principal hyperparameters to address. First is the initial learning rate, which controls the initial step size used by Adam to step down the gradient of the loss. The second hyperparameter is the mini-batch size, which sets the number of spectrograms to be passed through the model at one time. The optimal configuration is found through a grid search of the hyperparameters. A summary of the optimal hyperparameters and the number of spectrograms used are listed in Table 2.2. As seen in Figure 2.4a, training and validation losses fall off exponentially with each training epoch until the early stopping criterion is met; in this case, at 48 epochs. The effectiveness of the autoencoder's ability to reconstruct the input spectrogram is illustrated in Figure 2.5. Though some loss of resolution in time and frequency is expected due to the convolutional and transposed convolutional layers, the structure of the spectrogram is largely preserved, with the salient information of the input encoded to the latent feature space. To test that the autoencoder adequately generalized the entire data set, all spectrograms were fed through the model, yielding an average MSE of 5.9381×10^{-6}, which is consistent with the validation MSE at the early stopping point.

2.4.2 Clustering Methodologies

In our deep clustering framework, clustering is performed in the latent feature space, Z, to find 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 $\{\boldsymbol{\mu}_k \in Z\}_{k=1}^{K}$, i.e., centroids around which other similar signals may be found. We use Euclidean distance between a centroid and a latent feature vector to measure similarity:

$$d_{n,k} = \|\boldsymbol{z}_n - \boldsymbol{\mu}_k\|_2. \tag{2.3}$$

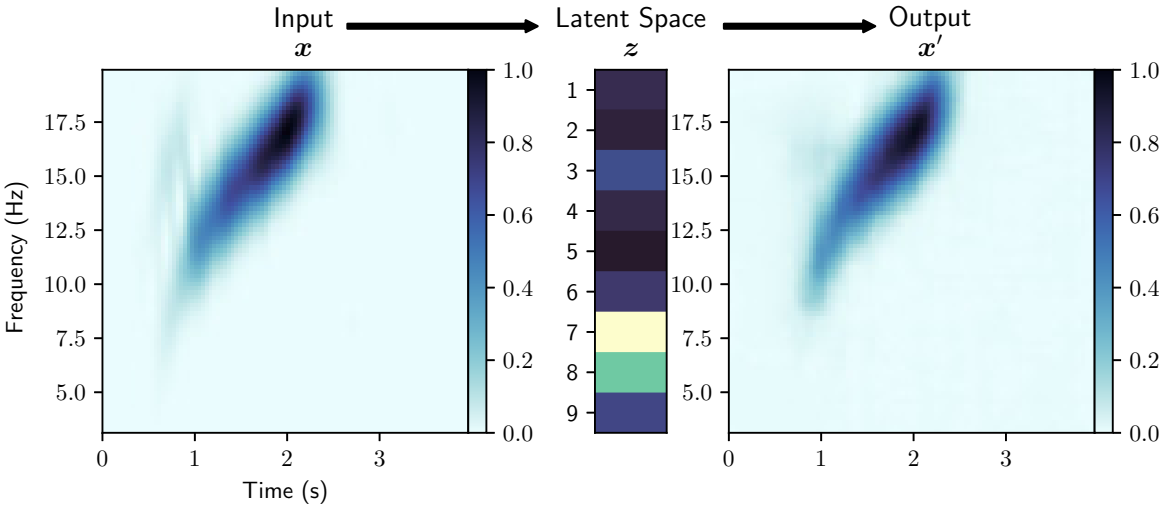

Figure 2.5. A trained autoencoder takes an input spectrogram $\boldsymbol{x}$, encodes it to a 9-dimensional latent feature vector $\boldsymbol{z}$, then reconstructs the input as $\boldsymbol{x}'$. The autoencoder preserves features correlated within a given cluster and discards the remaining signal, which can help with signal identification.

$d_{n,k}$ is a measure of the similarity between features indexed by n and k.

Gaussian Mixture Model (GMM)

In GMM clustering, the latent feature vectors $\boldsymbol{z}$ are described by a mixture of K Gaussian distributions that are linearly superimposed in the latent space Z, where each Gaussian model has its own centroid $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$. We follow the methods of [20, p. 430] and [57, p. 339]. The overall distribution of the mixture model is given by the convex combination of their distributions,

$$p(\boldsymbol{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{z} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{2.4}$$

Consider the latent feature vectors $\boldsymbol{z}_n$ as rows of a matrix $\mathbf{Z} \in \mathbb{R}^{N \times D}$ with N samples and D features. To estimate the parameters of each Gaussian distribution, an expectation-maximization (EM) algorithm is used to maximize the Gaussian mixture model's likelihood function of $\mathbf{Z}$ with respect to the parameters $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, and π_k [20, p. 433]:

$$\ln p(\mathbf{Z} \mid \{\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K\}, \{\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_K\}, \{\pi_1, ..., \pi_K\}) = \sum_{n=1}^{N} \ln \left[\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{z}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right]. \tag{2.5}$$

For every sample $\boldsymbol{z}_n$, a binary K-dimensional random variable $\xi_k \in \{0,1\}$ is introduced that has one element equal to one and all others to zero. The marginal distribution over $\boldsymbol{\xi}$ is $p(\xi_k = 1) = \pi_k$, where the mixing coefficients π_k satisfy $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$ in order to be valid probabilities. Since $\boldsymbol{\xi}$ is a 1-of-K (categorical) representation, this distribution is written as

$$p(\boldsymbol{\xi}) = \prod_{k=1}^{K} \pi_k^{\xi_k}, \tag{2.6}$$

and the conditional distribution of $\boldsymbol{z}_n$ given $\boldsymbol{\xi}$ as

$$p(\boldsymbol{z}_n \mid \boldsymbol{\xi}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{z}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{\xi_k}. \tag{2.7}$$

Equation (2.4) is then rewritten in terms of the factored joint distribution $p(\boldsymbol{z}_n, \boldsymbol{\xi}) = p(\boldsymbol{\xi})p(\boldsymbol{z}_n \mid \boldsymbol{\xi})$:

$$p(\boldsymbol{z}_n) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{z}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \sum_{\boldsymbol{\xi}} p(\boldsymbol{\xi})p(\boldsymbol{z}_n \mid \boldsymbol{\xi}). \tag{2.8}$$

Using Bayes' theorem and equations (2.4) and (2.8), the conditional probability of $\boldsymbol{\xi}$ given $\boldsymbol{z}_n$ is:

$$\gamma(\xi_k) \equiv p(\xi_k = 1 \mid \boldsymbol{z}_n) = \frac{p(\xi_k = 1)p(\boldsymbol{z}_n \mid \xi_k = 1)}{\sum_{j=1}^{K} p(\xi_j = 1)p(\boldsymbol{z}_n \mid \xi_j = 1)} = \frac{\pi_k \mathcal{N}(\boldsymbol{z}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{z}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}, \tag{2.9}$$

where π_k is the prior probability of $\xi_k = 1$, and $\gamma(\xi_k)$ is the posterior probability having observed $\boldsymbol{z}_n$. As with $\mathbf{Z}$, we construct a matrix $\boldsymbol{\Xi} \in \mathbb{R}^{N \times K}$ whose rows consist of the binary random variables $\boldsymbol{\xi}_n$ for each sample $\boldsymbol{z}_n$. Thus indexed, $\gamma(\xi_{nk})$ is defined as the *responsibility* that distribution k has for *explaining* sample $\boldsymbol{z}_n$, and is analogous to soft clustering, where the probability that sample $\boldsymbol{z}_n$ belongs to distribution k is determined for each of the K distributions. In practice, each latent feature vector $\boldsymbol{z}_n$ is assigned to one of K Gaussian distributions by $\arg\max_{\xi}[\gamma(\xi_{nk})]$.

Using superscript t to denote the iteration index, the EM algorithm for a Gaussian mixture

is:

1. Initialization of parameters $\boldsymbol{\mu}_k^{t-1}$, $\boldsymbol{\Sigma}_k^{t-1}$, and π_k^{t-1}.

2. Expectation step. This step encodes the samples' probability of assignment to each Gaussian distribution by evaluating responsibilities $\gamma(\xi_{nk})$ using $\boldsymbol{\mu}_k^{t-1}$, $\boldsymbol{\Sigma}_k^{t-1}$, and π_k^{t-1} (equation (2.9)).

3. Maximization step. Using the responsibilities $\gamma(\xi_{nk})$, this step updates the centroid location ($\boldsymbol{\mu}_k^t$), shape ($\boldsymbol{\Sigma}_k^t$), and normalization (π_k^t) of each distribution in the latent space Z by:

$$\begin{aligned}
\boldsymbol{\mu}_k^t &= \frac{1}{N_k}\sum_{n=1}^{N}\gamma(\xi_{nk})\boldsymbol{z}_n \\
\boldsymbol{\Sigma}_k^t &= \frac{1}{N_k}\sum_{n=1}^{N}\gamma(\xi_{nk})(\boldsymbol{z}_n-\boldsymbol{\mu}_k^t)(\boldsymbol{z}_n-\boldsymbol{\mu}_k^t)^{\mathrm{T}} \\
\pi_k^t &= \frac{N_k}{N}
\end{aligned} \tag{2.10}$$

where

$$N_k = \sum_{n=1}^{N}\gamma(\xi_{nk}).$$

4. Convergence check. The log likelihood of $\mathbf{Z}$ is evaluated with respect to the parameters $\boldsymbol{\mu}_k^t$, $\boldsymbol{\Sigma}_k^t$, and π_k^t (equation 2.5). If convergence occurs in the log likelihood or in the parameters $\boldsymbol{\mu}_k^t$, $\boldsymbol{\Sigma}_k^t$, and π_k^t, the EM algorithm has reached a local maximum and terminates; otherwise, the algorithm returns to step 2.

To accelerate EM convergence, k-means clustering is used to initialize the GMM clustering algorithm [20, p. 438]. EM stops after 1,000 iterations have elapsed or when the change in log likelihood from equation (2.5) is less than 0.001. To avoid converging on local maxima, the initialization is run 100 times and the initialization with the best log likelihood is retained.

Deep Embedded Clustering (DEC)

In DEC, clustering is performed in conjunction with continued training of the autoencoder, with the clustering layer attached to the bottleneck providing an additional loss function that is backpropagated through the autoencoder layers (Figure 2.3). The DEC model DNN parameters

are initialized using the parameters of the trained autoencoder, and clustering layer parameters are initialized using the centroids from GMM clustering. DEC seeks to improve the GMM clustering by using the Euclidean distance between embedded spectrograms and cluster centroids (equation (2.3)) as an additional loss function for updating model parameters. Because the input data is unlabeled, a self-supervised method is required. We implement the method developed by [64], who, drawing from the t-distributed stochastic neighbor embedding (t-SNE) algorithm [75], propose measuring the difference between a Student's t-distribution kernel of the latent feature vectors $\boldsymbol{z}$ and an auxiliary target distribution. A simplified Student's t-distribution is used to measure the similarity between embedded spectrograms $\boldsymbol{z}_n$ and the cluster centroids $\boldsymbol{\mu}_k$:

$$q_{nk} = \frac{(1+ \parallel \boldsymbol{z}_n - \boldsymbol{\mu}_k \parallel^2)^{-1}}{\sum_k (1+ \parallel \boldsymbol{z}_n - \boldsymbol{\mu}_k \parallel^2)^{-1}}. \tag{2.11}$$

Equation (2.11) results in a set of soft class assignments, i.e., the probability that embedded spectrogram n will be assigned to class k. Latent feature vectors $\boldsymbol{z}_n$ are assigned to one of K classes by $\arg\max_q [q_{nk}]$. The soft class assignments q_{nk} are then used to compute the auxiliary target distribution, p, whose form 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 [64]:

$$p_{nk} = \frac{q_{nk}^2 / \sum_n q_{nk}}{\sum_k (q_{nk}^2 / \sum_n q_{nk})}. \tag{2.12}$$

The dissimilarity between the distributions given by equations (2.11) and (2.12) is measured using the Kullback-Leibler divergence [76]. From the divergence the clustering layer's loss function is obtained:

$$L_{\mathrm{C}} = D_{\mathrm{KL}}(P \parallel Q) = \sum_n \sum_k p_{nk} \log \frac{p_{nk}}{q_{nk}}. \tag{2.13}$$

In DEC, the clustering layer is attached to the trained autoencoder's bottleneck. During training of the DEC model, the loss functions from equations (2.2) and (2.13) are combined into

a total loss function,

$$L = L_{\mathrm{AEC}} + \lambda L_{\mathrm{C}}, \tag{2.14}$$

where λ is a hyperparameter 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 $\lambda = 10^{-4}$ yielded optimal performance for model training and clustering.

Two constituent processes occur simultaneously during DEC model training. First, the full loss from equation (2.14) 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_{nk} and p_{nk} are updated at intervals. The update interval is a hyperparameter that must be tuned. Through hyperparameter tuning, an update interval of 10 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 0.4% of the total number of training samples. The same mini-batch size and initial learning rate are used to train both the autoencoder and DEC model (Table 2.2). Figure 2.4b shows how losses decrease over time and the percent change in label assignments for every mini-batch training iteration. Though the overall trends in the loss curves show exponential decay, periodic spikes occur at every update interval, when q_{nk} and p_{nk} are recalculated, and are visible since the losses are recorded after every mini-batch rather than every epoch.

2.4.3 Selecting Optimal Number of Clusters

Determining the optimal number of clusters, K, is a major challenge in unsupervised machine learning. In this study we treat K as a hyperparameter, iterating the deep clustering

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, $d_{n,k}$ (equation (2.3)). Additionally, traditional statistical methods for choosing the optimal number of clusters, such as the gap statistic [77] and silhouette score [78], are consulted. The qualitative approach is to visually inspect the similarity of the latent feature vectors $\boldsymbol{z}_n$ to their respective class centroids $\boldsymbol{\mu}_k$, 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 in those classes 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 RIS data set.

2.5 Results

The following analysis of GMM and DEC performance focuses on how the clustering algorithms affect the latent space Z and whether the methods yield meaningful results in the data space X. Since the samples in the data set are unlabeled and there is no "ground truth" against which to compare results, measurements of intra-class similarity among spectrograms and latent feature vectors are examined. We conclude that neither GMM nor DEC provides a clear advantage in clustering performance. Accordingly, we recommend implementation of GMM for deep clustering of RIS seismic data. The statistical and mathematical underpinnings of GMM are well understood, and the complexity of implementation and interpretation of DEC is difficult to justify in the absence of compelling performance improvement. Furthermore, in practice GMM clustering on a graphics processing unit takes approximately one minute to cluster the entire data set, whereas one DEC hyperparameter tuning run can take several hours.

In the analyses that follow, results are presented for the entire data set of 531,407

spectrograms, including the training and validation data subsets. We mitigate the risk of the DNN in the DEC model overfitting on the training data [57, p. 23] by using less than 10% of the data set for training and validation, and by drawing training samples randomly without replacement to achieve a training subset representative of the entire data set.

2.5.1 Clustering Performance

Deep clustering performance is qualitatively checked by comparing centroids to their respective assigned latent data samples. Results for GMM are shown in Figure 2.6. Each class k is represented by the columns in Figure 2.6, with each centroid $\boldsymbol{\mu}_k$ and its reconstruction $g_\theta(\boldsymbol{\mu}_k)$ 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 $\boldsymbol{x}_n$ assigned to its class. Subsequent rows show the latent feature vectors $\boldsymbol{z}_n$, spectrograms $\boldsymbol{x}_n$, and associated seismograms of the data samples assigned to the respective classes. To inspect whether intra-class similarity holds with increasing distance from the centroid, samples $\boldsymbol{z}_n$ and $\boldsymbol{x}_n$ are shown for $n = \{1, 1000, 5000, 10000, 15000, 20000, 25000\}$. Near the centroid, latent feature vectors $\boldsymbol{z}_n$ generally exhibit similar values to their class centroid $\boldsymbol{\mu}_k$, indicating that GMM has successfully grouped similar latent data samples into the 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 $g_\theta(\boldsymbol{\mu}_k)$, 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 time-frequency domain extends to the time domain, where seismograms in each class are similar to one another. As distance increases (i.e., with increasing n), cases of dissimilarity begin to arise as samples overlap with adjacent clusters.

In addition to checking the efficacy of the clustering, visual examination of the results in Figure 2.6 gives 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 amplitude characteristics. If such distinctions are not useful or if

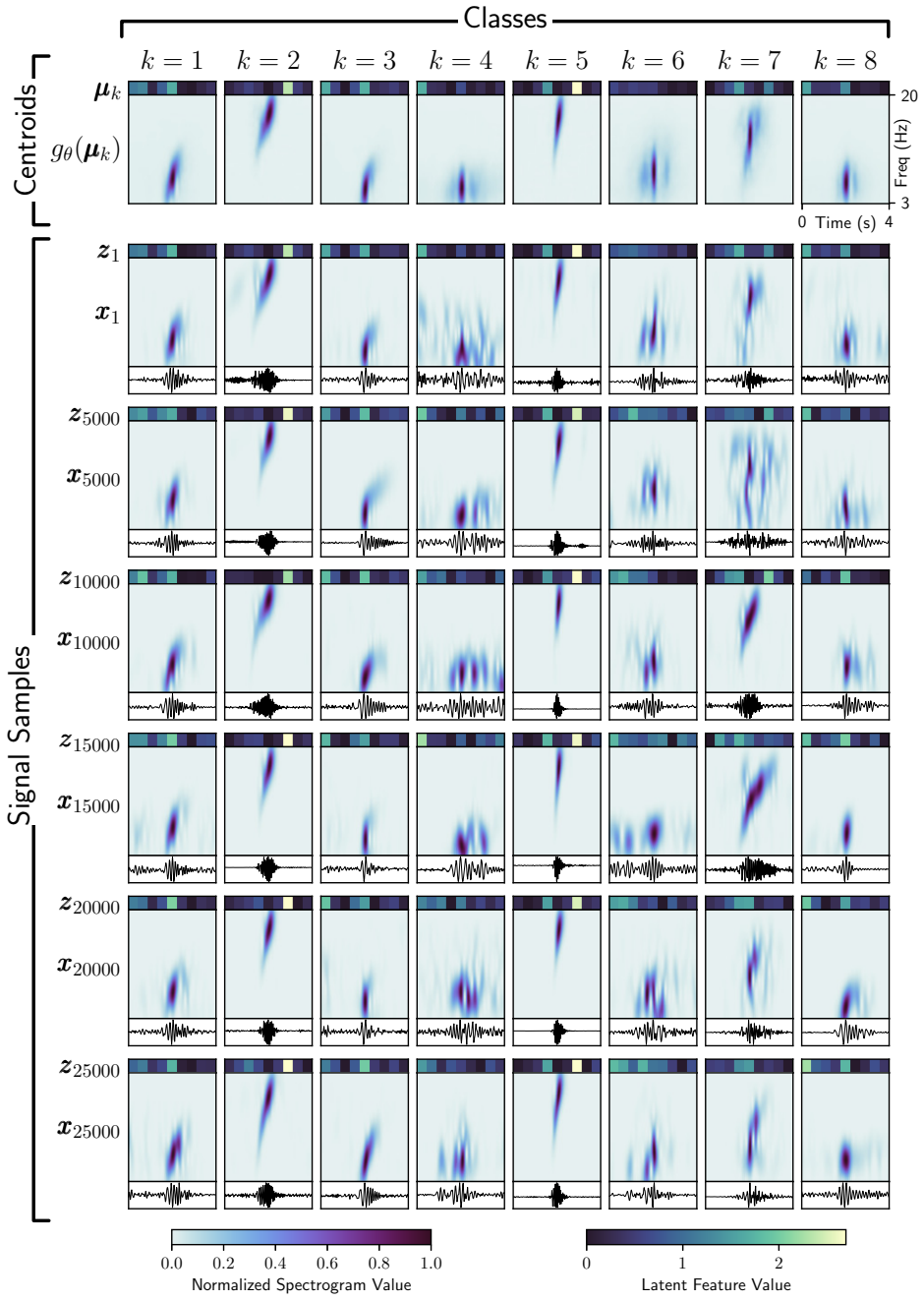

Figure 2.6. Gaussian mixture model (GMM) clustering results are shown, with samples $\boldsymbol{z}_n$ and $\boldsymbol{x}_n$ the n^{th} closest to their respective centroids. Within a given class k, the cluster centroids $\boldsymbol{\mu}_k$ are similar to the latent feature vectors $\boldsymbol{z}_n$, whose nine elements are shown above each spectrogram. Though the centroids are not members of the data set, their reconstructions $g_\theta(\boldsymbol{\mu}_k)$ exhibit similar characteristics to the spectrograms $\boldsymbol{x}_n$ assigned to each class. Seismograms plotted below each spectrogram also exhibit similarity within each class. With increasing distance from the centroid (i.e., as n increases), dissimilarity and potential cases of mis-assignment are visible in latent feature vectors, spectrograms, and seismograms, e.g for $k = 7$, $n = 15000$.

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.

Clustering with DEC involves two steps: first, the GMM clustering algorithm initializes the 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. To determine to what extent this occurs, t-SNE is used to visualize the 9-dimensional latent space in two dimensions [75]. t-SNE can illuminate possible clusters within data in an unsupervised manner by displaying data in geometrically separated clusters. In Figure 2.7a, t-SNE results of the latent feature space clustered with GMM show that the data are largely contiguous with few exceptions. Applying the labels assigned by GMM clustering 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. Contrast this with Figure 2.7b, in which t-SNE results at the conclusion of DEC show both geometric separation as well as nearly 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 deep clustering on the densities of the clusters provides a more concrete visualization. Of interest to the ability for the clustering algorithms to identify clusters is the distance of each cluster to the others. In Figure 2.7c, the probability density functions (PDF) of all clusters are shown as functions of distance to each centroid. Before DEC, though GMM clustering usually 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. With 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

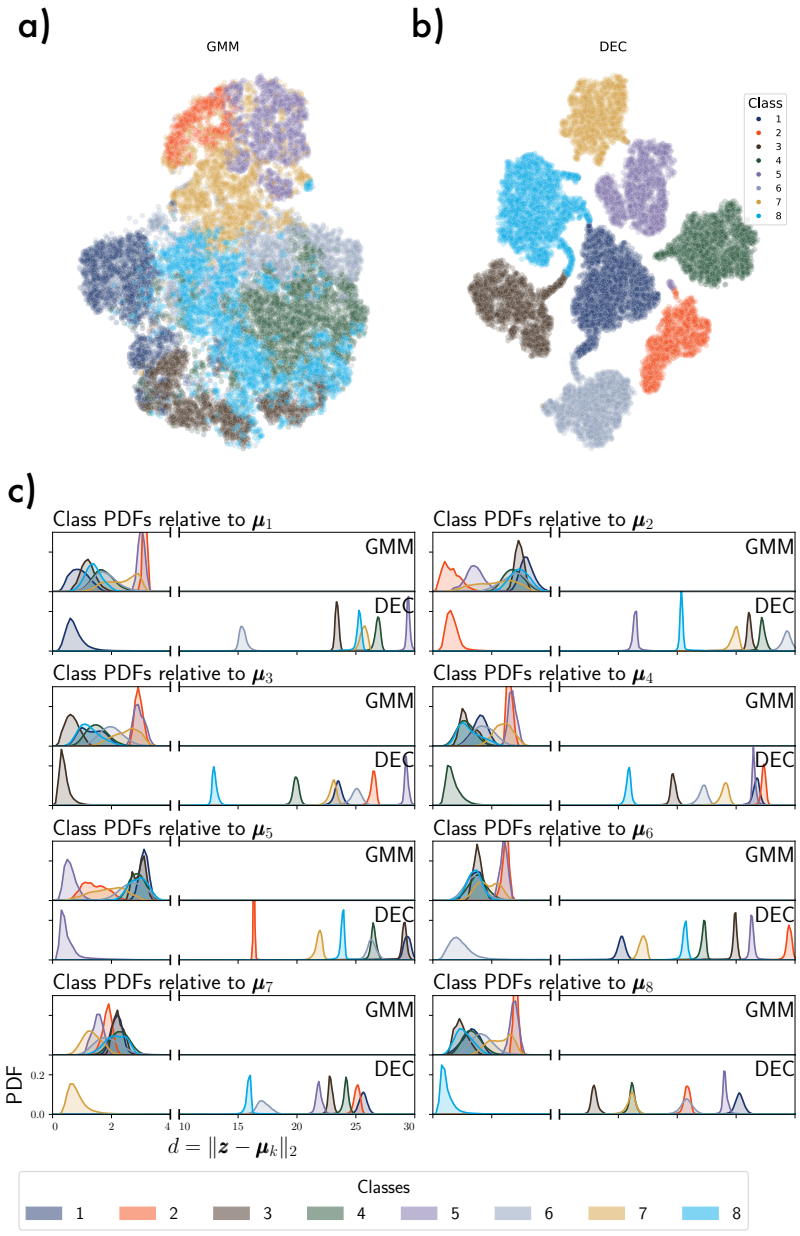

Figure 2.7. (a) Visualization of the 9-dimensional latent data space is shown in two dimensions using the t-distributed stochastic neighbor embedding (t-SNE) plot for Gaussian mixture model (GMM) clustering. GMM exhibits limited separation within the data and overlapping classes. (b) t-SNE plot for deep embedded clustering (DEC), whose clusters are well separated and contain nearly homogeneous class members. (c) The effects of DEC in the latent feature space are 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 increases the distance between the other class centroids and PDFs.

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 2.8. By sorting the latent space by sample index n such that $d_{n+1,k} > d_{n,k}$, cluster separation can be visualized directly in the latent space. Before DEC, centroids are initialized with the GMM clustering 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 n increases. The contrast with the latent feature 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. With GMM, 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.

While DEC effectively transforms the latent feature space Z by moving latent feature vectors closer to their centroids, less clear is whether this transformation causes a corresponding improvement in clustering quality in the data space X. To evaluate intra-class similarity among spectrograms, four pairwise metrics are used to compare the clustering assignments obtained from GMM and DEC.

The first metric used is the silhouette coefficient, which uses the mean intra-cluster and nearest-cluster distances to express whether a sample belongs in its assigned cluster or if it is more similar to another cluster [78]. The silhouette coefficient exists on the interval $[-1,1]$, with positive values indicating a sample has likely been correctly assigned, values near 0 indicating overlapping clusters, and negative values indicating a sample may have been placed in the wrong cluster. Coefficients are calculated for every sample, and the silhouette score is defined as the mean of all the coefficients. A summary of class and total silhouette scores is given in Table 2.3.

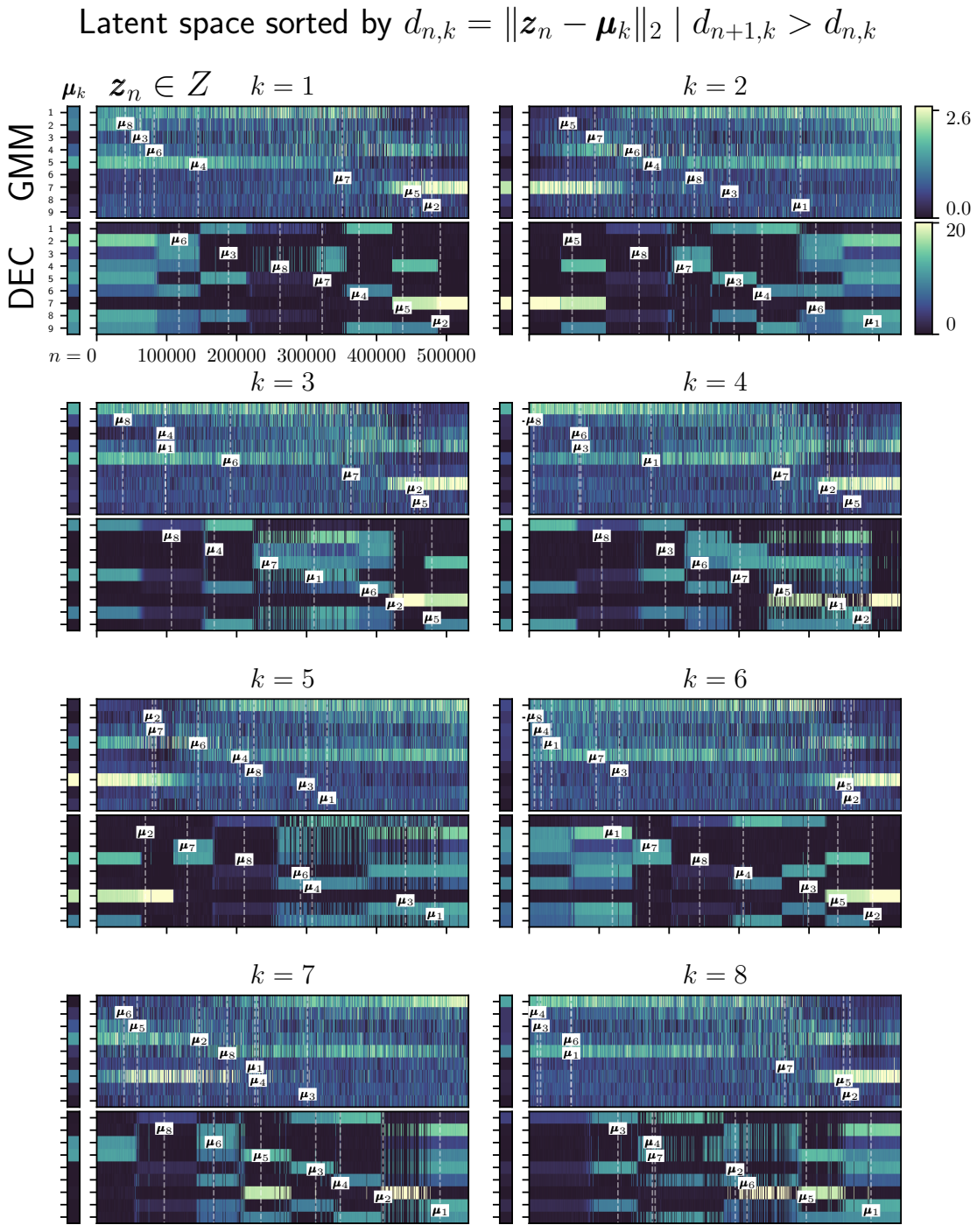

Figure 2.8. For each class k, latent data samples $\boldsymbol{z}_n$ are shown stacked according to their distance $\|\boldsymbol{z}_n - \boldsymbol{\mu}_k\|$ from the centroid $\boldsymbol{\mu}_k$ (shown to the left). Distance of the other cluster centroids relative to the selected class k are indicated with vertical dotted lines. Deep embedded clustering (DEC) brings assigned data $\boldsymbol{z}_n$ closer to the class centroid, resulting in homogeneity among the latent feature vectors assigned to that class.

In Figure 2.9, silhouette analyses are shown stacked by cluster assignment for the latent feature data in Z for GMM (Figure 2.9a) and DEC (Figure 2.9b), and for the spectrograms in the data space X for GMM (Figure 2.9c) and DEC (Figure 2.9d). In Figure 2.9a, classes 1-3 and 5 are decently clustered, classes 4, 6, and 7 are likely in a region of overlap, and class 8 is not well clustered; the silhouette score for this data is 0.08. In contrast, every class in Figure 2.9b is well clustered with a silhouette score of 0.90, results which are consistent with those presented in Figures 2.7 and 2.8. To determine whether these analyses correspond to meaningful results in the data space, we examine the correlation between the silhouette analyses of the latent space Z and data space X. The silhouette analysis for GMM in the data space is shown in Figure 2.9c with a silhouette score of 0.05. These results are consistent with the GMM latent space results in Figure 2.9a and indicate a proper mapping from the data space into the latent space with the autoencoder. The silhouette analysis for DEC in the data space is shown in Figure 2.9d with a silhouette score of 0.13, which is inconsistent with its corresponding latent space analysis in Figure 2.9b. Comparison between Figures 2.9c and Figures 2.9d might lead us to conclude that DEC provides superior clustering performance, and this may be true. However, the inconsistency observed for DEC between the latent space and the data space require that additional metrics be examined.

For the remaining metrics, spectrograms $\boldsymbol{x}_n$ are vectorized and divided by their vector norm, resulting in unit vectors projected onto an n-sphere. The second metric is obtained by taking the inner product between two such unit vectors, which provides a measure of the angle between them and thus a proxy for similarity. The third metric is MSE, but to mitigate its tendency to exaggerate the effects of outliers by squaring the error, the mean absolute error (MAE) is used as a fourth metric. For each of these metrics, an intra-class mean vector is calculated against which all other vectors in the class are measured. The class and total mean values for each metric for GMM and DEC are given in Table 2.3, with better scores in bold. While Figures 2.7, 2.8, and 2.9b,d may lead us to favor DEC performance, the data space metrics in Table 2.3 offer a more nuanced understanding. On average, DEC slightly outperforms GMM in the mean inner

Table 2.3. Comparison of Clustering Metrics for Gaussian Mixture Model (GMM) Clustering and Deep Embedded Clustering (DEC)

Class	N	Data Space				Latent Space
		Mean Inner Product	Mean MSE ($\times 10^{-5}$)	Mean MAE ($\times 10^{-3}$)	Silhouette Score	Silhouette Score
1	66817 / 85789	**0.82** / 0.80	**0.26** / 0.28	**0.20** / 0.23	0.19 / **0.20**	0.11 / **0.89**
2	27568 / 45607	**0.88** / 0.81	**0.44** / 0.55	**0.36** / 0.44	**0.31** / 0.20	0.39 / **0.93**
3	59131 / 63725	0.86 / **0.87**	**0.64** / 0.74	**0.53** / 0.61	**0.27** / 0.26	0.30 / **0.90**
4	95323 / 68521	0.61 / **0.73**	1.21 / **1.13**	0.90 / **0.90**	-0.08 / **0.11**	0.00 / **0.92**
5	57318 / 64235	**0.91** / 0.85	**1.33** / 1.35	**1.01** / 1.05	**0.41** / 0.30	0.41 / **0.93**
6	63326 / 59925	0.49 / **0.64**	2.06 / **1.87**	1.48 / **1.43**	-0.10 / **-0.03**	-0.08 / **0.85**
7	61430 / 55699	0.48 / **0.57**	2.81 / **2.49**	1.88 / **1.82**	-0.09 / **-0.08**	-0.08 / **0.89**
8	98494 / 87906	0.67 / **0.76**	3.29 / **2.84**	2.16 / **2.10**	-0.14 / **-0.01**	-0.08 / **0.87**
	Overall Mean:	0.71* / **0.75***	1.50* / **1.41***	**1.06*** / 1.07*	0.05 / **0.13**	0.08 / **0.90**

All table values read as GMM / DEC. *Weighted mean.

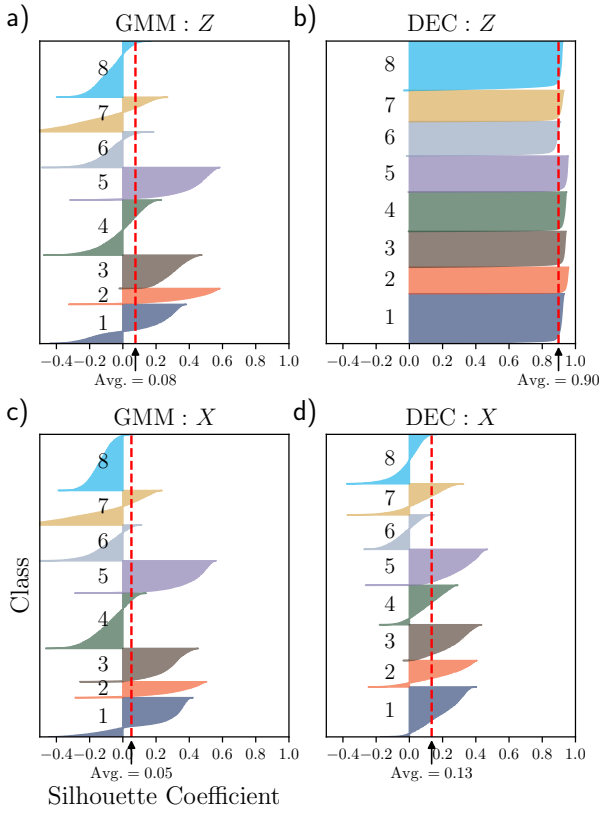

Figure 2.9. Silhouette analyses for (a,c) Gaussian mixture model (GMM) clustering and (b,d) deep embedded clustering (DEC) for the (a, b) latent feature space Z and (c,d) data space X.

product, MSE, and silhouette score. Importantly, however, the inconsistencies among the metrics within each class preclude a definitive decision regarding which clustering method is better. Of particular concern is the disparity in latent space and data space results for DEC. The latent space transformation in DEC is substantial and does lead to sharp, distinct clusters in the latent space. However, it appears these results do not map into the data space so readily. We assess that this disparity arises when the DEC model is training: as the model parameters are updated, the latent space is continually manipulated to conform to the class centroids, effectively distorting the latent space. Even through hyperparameter tuning, we were unable to obtain results that provided a compelling reason to justify the complexity of DEC, especially within the context of initial data exploration, in which GMM is more efficient. Consequently, results shown in the subsequent sections are from the GMM deep clustering workflow.

2.5.2 Deep Clustering Methodology Considerations

One of the key strengths of the deep clustering implementation in this study is the 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 and types 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.

The selection of an appropriate algorithm for the clustering layer largely depends on the type and properties of the data set. Though in this study we use GMM and DEC, as described in Section 2, there are numerous clustering algorithms of which some may be applicable to a deep clustering workflow. Regardless of the choice of clustering algorithm, careful consideration

must be given towards understanding whether clustering in the latent space maps to meaningful results in the data space.

The flexibility afforded by deep clustering 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 deep clustering workflows [39]. In our case, we used spectrograms, but other transforms, such as continuous wavelet transform scalograms, could just as easily be used as inputs. In post-processing, redundant or similar results can be combined.

2.6 Discussion: Glaciological Implications

The spatial and temporal distribution of signals from the eight classes identified gives information on the response of the RIS to various climatological forcings, including from oceanographic and atmospheric variability. Importantly, two years of continuous seismic monitoring allows identification of seasonal and interannual patterns of variability, particularly allowing examination of the effects of the strong 2016 El Niño on RIS seismicity by comparisons with 2015 levels.

The two-year RIS array data set contains 531,407 seismic detections. A summary of the data set statistics and class characteristics (Table 2.4) 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 2, 4, 5, 6, and 8 have pronounced differences (more than 10%) between the number of detections occurring in the summers versus the winters, while differences for classes 1, 3, and 7 are less pronounced (between 5% and 10%). Interannual comparisons for each season show that classes 5, 6, and 7

Table 2.4. 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

Class	Detections			Freq (Hz)	Amplitude (accel., nm/s^2)			
	N	%N Summer (JFM) Total | 2015 | 2016	%N Winter (JJA) Total | 2015 | 2016	Mean peak	Mean	Median	Std. dev.	Max.
1	66,817	27 | 13 | 13	22 | 11 | 11	7.3	46	37	45	3,242
2	27,568	1 | 0 | 1	27 | 0 | 27	16.7	60	27	95	2,222
3	59,131	30 | 16 | 14	21 | 11 | 10	5.9	61	37	130	12,825
4	95,323	37 | 17 | 20	23 | 10 | 13	5.4	112	32	488	41,924
5	57,318	13 | 0 | 12	29 | 1 | 28	16.6	124	42	368	33,623
6	63,326	39 | 16 | 23	19 | 8 | 11	8.1	155	34	6,533	1,632,100
7	61,430	24 | 6 | 18	19 | 3 | 16	13.7	169	30	3,277	461,205
8	98,494	46 | 22 | 24	16 | 7 | 9	6.3	210	46	1,388	268,633

experienced an increase in activity in the 2016 austral summer over the 2015 austral summer, with classes 5 and 7 exhibiting the largest changes.

The seasonal changes are investigated in more detail in Figure 2.10a, where detection occurrences shown as a function of station and month exhibit spatiotemporal patterns that reveal associations between environmental forcing and seismicity. Clustering enables these patterns to be further explored by class and month (Figure 2.10b), and by class and station (Figure 2.10c).

From Figure 2.10a, 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 (Figure 2.1) and detected seismicity associated with ocean gravity waves impacting the shelf front that cause fracturing (icequakes) and calving [35]. Seismicity at these stations during the 2016 austral summer was higher than the same period in 2015, and across the array, a substantial increase in seismicity was observed in the months immediately following the 2016 austral summer, indicative of the impact of El Niño on Antarctic ice shelf fronts [79].

Some of the most seismically active stations were located near grounding zones: station RS09 (118,105 detections) on the eastern flank of Roosevelt Island; station RS11 (81,138 detections) on the Shirase Coast; station RS17 (50,385 detections) on Steershead Ice Rise; and

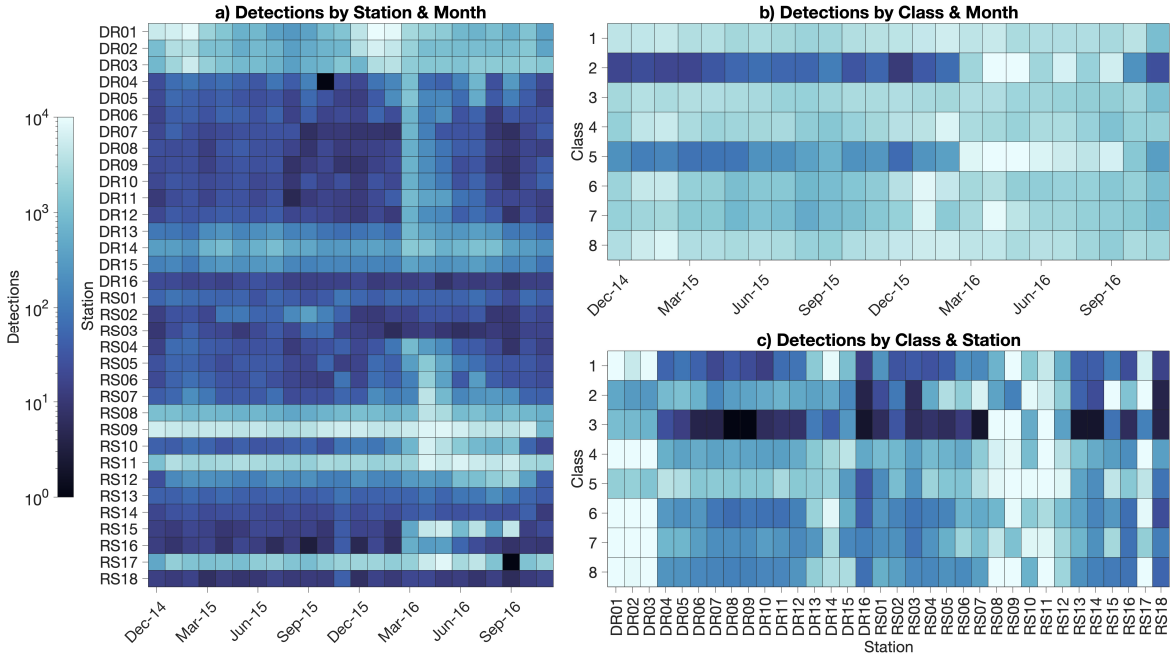

Figure 2.10. (a) The frequency of detections comprising the Ross Ice Shelf data set is shown by station and month. Clustering provides a further breakdown by (b) class and month for all stations, and (c) class and station.

station RS08 (25,500 detections) on the western flank of Roosevelt Island. These stations were on either fully or partially grounded ice, suggesting that the seismicity results from interactions of basal ice with the solid earth. Increases in seismicity during the 2016 winter at floating stations RS10 (between Roosevelt Island and the Shirase Coast) and RS15 (over a bathymetric high) may result from El Niño related changes in water layer thickness that affect flexural gravity wave amplitudes [32]. The RIS front stations DR01 (64,311 detections), DR02 (39,822 detections), and DR03 (39,176 detections) were also active. All of these active 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 2.10b. Classes 2 and 5 have increased detection frequencies in the austral winter of 2015 when local storms are more intense, suggesting meteorological forcing. The remaining classes have increased detections in the austral summers. The clustering results reveal that the large increase in seismicity

in classes 2 and 5 occurs following the 2016 austral summer. A further dimension to the analysis is shown in Figure 2.10c, which shows the distribution of classes by station. Classes 1, 4, 5, and 8 are prominent signal types at stations near grounding zones (RS08, RS09, RS11, and RS17), and classes 1, 4, 6, 7, and 8 are prominent at the RIS front (DR01, DR02, DR03).

An important caveat for the detection statistics shown in Table 2.4 and Figure 2.10 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 originating farther away than high-frequency signals. For example, from Figure 2.6, class 1 is similar to class 3, with the notable difference in that class 1 contains more energy at frequencies slightly higher than class 3 and has lower amplitude. Thus, class 3 may be generated by a similar source mechanism as class 1 but have a longer propagation path.

Factoring in signal amplitude also affects the range at which seismic energy is detected. From Table 2.4, class 2 has an average spectral peak at 16.7 Hz, the highest of the classes, with a total of 27,568 detections, the lowest of the classes. Similarly, class 5 has the second-highest average spectral peak at 16.6 Hz, with the second lowest amount of detections among the classes. These two classes are nevertheless distinct from each other in amplitude and waveform type: from Table 2.4, class 2 has a mean amplitude of 46 $\mathrm{nm/s^2}$, while class 5 has a mean amplitude of 124 $\mathrm{nm/s^2}$. From Figure 2.6, class 2 consists of high frequency signals experiencing dispersion, while class 5 signals are more impulsive; both likely result from fracturing.

Detection statistics are 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 nearly simultaneously. Consideration should also be given to determining if classes are duplicates of the same seismic source mechanism. Seismic surface waves in the ice undergo dispersion as they propagate, which DEC may interpret as separate signal classes. This may be be the case with classes 2 and 5. The longer wave train for class 5 signals is consistent with Rayleigh wave propagation of class 2 signals. Propagation

modeling can be used to calculate expected dispersion relations to confirm if this is the case. Such distinctions could be useful in identifying common propagation paths or providing source range discrimination.

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 deep clustering in exploring data and identifying potential causes of seismicity when examined in conjunction with environmental data.

2.6.1 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 winter, sea ice coverage (Figure 2.11a) reaches nearly 100%, damping ocean swell. During the austral summer, sea ice concentration decreases to approximately 25%, permitting ocean gravity waves to directly impact the ice shelf front and cause iceberg calving. Additionally, warmer air temperatures (Figure 2.11b) may promote calving with associated increased icequake activity [35].

Increased levels of seismicity at DR02 are observed for all classes except 2 and 5 at DR02 (Figure 2.11d,f,g,i-k) during the austral summers. Classes 4, 6, and 8 are especially active during the 2016 austral summer, when strong El Niño conditions led to anomalously persistent high temperatures across West Antarctica [79] and ocean-ice shelf interactions were 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 2.10a. Widespread surface melt on the RIS was observed between 10-21 January 2016 [79, 30], which affects firn layer properties and seismicity through freeze/thaw cycles [29].

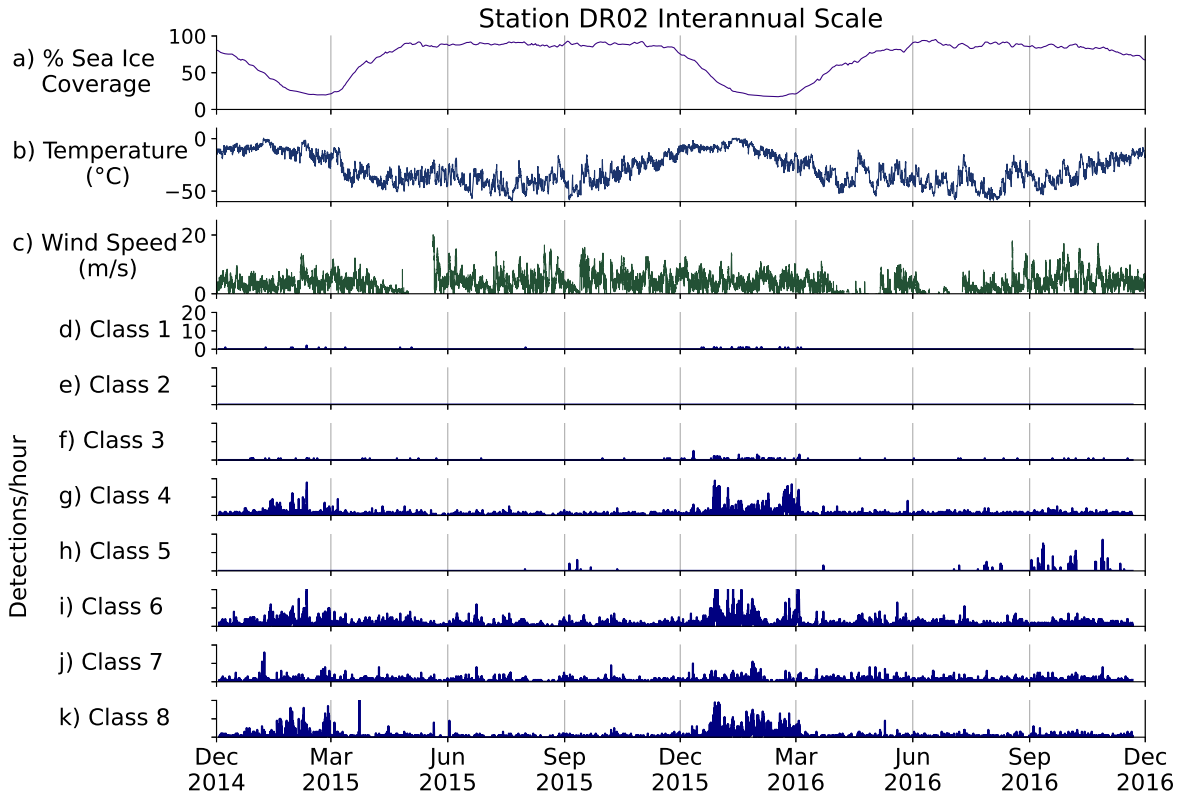

Figure 2.11. Two years of (a) sea ice coverage on the Ross Sea, (b) temperature and (c) wind speed observations at Gill automated weather station (approximately 223 km south of DR02, Figure 2.1), and (d-k) icequake detection statistics for each signal class. Classes 4, 6, 7, and 8 exhibit increased seismicity during the austral summers. Sea ice concentration data were obtained from NSIDC [4]; weather station data from AMRC, SSEC, UW–Madison.

Although class 6 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 [80], as the ice flow velocity in the vicinity of station DR02 is among the highest observed on the RIS. Class 5 (Figure 2.11h) is more active during the coldest periods of the year (April-September), suggesting that these signals may be associated with extremely cold temperatures or strong wind events. Cold-weather enhanced seismicity occurs at a rift approximately 140 km south of the ice front [28]. Alternatively, from Table 2.4, these classes are lower amplitude than those most active during the austral summer, which suggests that these detections may be masked by higher amplitude signals associated with the other classes. 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.

2.6.2 Diurnal seismicity on Roosevelt Island

Station RS09 on the eastern flank of Roosevelt Island experienced the most detections across the array, comprising 22% of detections in the full data set. In Figure 2.12, potential environmental sources of seismicity are compared to the seismicity of each class. Temperature and wind speed (Figure 2.12a,b) were recorded at a nearby automated weather station, Margaret, 122 km southwest of RS09. Tides (Figure 2.12c) were realized from the CATS2008 model [5] 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 class 1 (Figure 2.12d) dominates the detections at RS09 and is active throughout the year, comprising 52.8% of the detections. Classes 3, 4, 6, and 8 (Figure 2.12f,g,i,k) are also active throughout the year. Classes 5 and 7 (Figure 2.12h,j) are comparatively sparse, with seismicity limited to what appear to be discrete signals that could be associated with large fracture or crevasse events. No class 2 (Figure 2.12e) signals were recorded at RS09, even though elevated class 5 seismicity occurred during the 2016

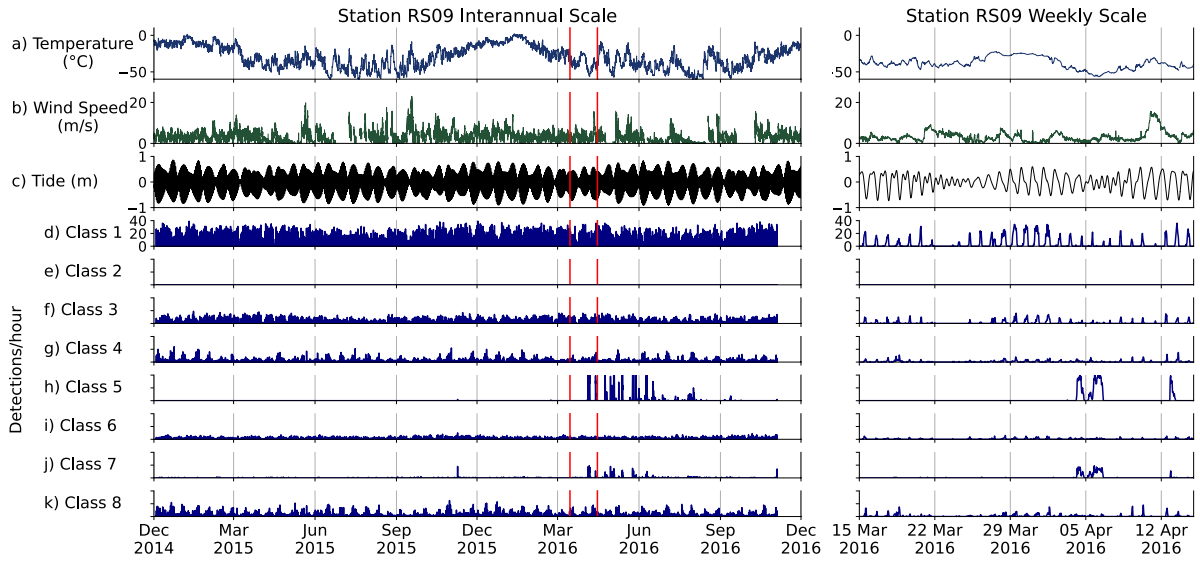

Figure 2.12. Two years of (a) temperature and (b) wind speed observations at Margaret automated weather station (MGT, approximately 122 km southwest of RS09, Figure 2.1), c) model-derived tides calculated at station RS10, and (d-k) icequake detection statistics for each signal class. Interannual timescale is shown at left with vertical red lines indicating the subset weekly timescale at right. The diurnal tidal signal correlates with seismicity for classes 1, 3, 4, 6, and 8. Tidal model from [5]; weather station data from AMRC, SSEC, UW–Madison.

winter.

Of particular interest at station RS09 is evidence of seismicity associated with the diurnal tide (Figure 2.12). On an interannual timescale, classes 4 and 8 exhibit a periodic modulation of seismicity which tends to correlate with spring tides. Variability over fortnight tidal cycles is shown between 15 March 2016 and 15 April 2016. This weekly timescale shows that classes 1 and 3 correlate with diurnal tides. Even some relatively non-active classes (4, 6, 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 [55]. The weekly timescale also reveals the sudden onset and termination of winter seismicity in classes 5 and 7, suggesting association with discrete ice shelf events such as crevasse expansion or major ice fracture. This onset is consistent with the substantial increase in seismicity detected across the RIS array visible in Figure 2.10 beginning in March 2016.

Other stations located at grounding zones exhibit similar patterns of seismicity, though

to a lesser extent than RS09. Station RS11, located east of RS09 on the Shirase Coast, exhibits patterns of seismicity similar to RS09. These similarities indicate that ice shelf seismicity at grounding zones is associated with similar ice shelf processes. RS08, on the western flank of Roosevelt Island, and RS17, at Steershead Ice Rise, also exhibit diurnal seismicity, suggesting a dynamic diurnal process common to the grounding zones. These patterns of seismicity indicate that the interaction of the ice shelf with the solid earth at grounding zones is modulated by tides. Among the four stations at grounding zones, classes 1, 4, and 8 are the most common signals, with class 8 signals occurring most frequently at these stations. With a mean peak frequency of 6.3 Hz and a mean amplitude of 210 nm/s^2, class 8 signals are among the strongest detected across the array.

2.7 Conclusions

Deep clustering of the Ross Ice Shelf (RIS) seismic array data set using a Gaussian mixture model identified eight classes of impulsive signals, with linkage of at least two of the classes to tidal variability near grounding zones. Additionally, compared to 2015, stations near the RIS front showed increased icequake activity during the 2016 El Niño austral summer. A sudden increase in seismicity was also observed across the array during the transition to the 2016 austral winter. The highest seismicity was observed at grounding zones, particularly along the eastern flank of Roosevelt Island.

Deep clustering is an effective way to explore large seismic data sets, particularly in its ability to identify dominant types of seismicity. The results provided by deep clustering, 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, deep clustering 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 deep clustering can be incorporated into existing seismic workflows to speed up exploratory

data analysis.

As seismic data sets grow ever larger, novel machine learning techniques will be necessary to enable researchers to fully utilize this data. Deep clustering 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.

2.8 Acknowledgements

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 (NSF) grant PLR 1246151, with support for Bromirski by NSF 1744856. Seismic data from network XH [81] were downloaded through IRIS Web Services (https://service.iris.edu/irisws/). Seismic data were processed using Obspy software [82]. 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 [2] and plotted using Antarctic Mapping Tools for MATLAB [3]. Surface temperatures were obtained from AMRC, SSEC, University of Wisconsin–Madison (https://amrc.ssec.wisc.edu). Tide data were generated by the CATS2008 model [5]. Ross Sea ice coverage was obtained from NASA NSIDC [4]. Code for this workflow is available at https://github.com/NeptuneProjects/RISClusterPT.

Chapter 2, in full, is a reprint of the material as it appears in the Journal of Geophysical Research: Solid Earth 2021. Jenkins, William; Gerstoft, Peter; Bianco, Michael; Bromirski, Peter, American Geophysical Union, 2021. The dissertation author was the primary investigator and author of this paper.

2.9 References

[1] P. D. Bromirski, A. Diez, P. Gerstoft, R. A. Stephen, T. Bolmer, D. A. Wiens, R. C. Aster, and A. Nyblade, "Ross ice shelf vibrations," *Geophysical Research Letters*, vol. 42, pp. 7589–7597, Sept. 2015.

[2] M. Morlighem, C. N. Williams, E. Rignot, L. An, J. E. Arndt, J. L. Bamber, G. Catania, N. Chauché, J. A. Dowdeswell, B. Dorschel, I. Fenty, K. Hogan, I. Howat, A. Hubbard, M. Jakobsson, T. M. Jordan, K. K. Kjeldsen, R. Millan, L. Mayer, J. Mouginot, B. P. Y. Noël, C. O'Cofaigh, S. Palmer, S. Rysgaard, H. Seroussi, M. J. Siegert, P. Slabon, F. Straneo, M. R. van den Broeke, W. Weinrebe, M. Wood, and K. B. Zinglersen, "BedMachine v3: Complete Bed Topography and Ocean Bathymetry Mapping of Greenland From Multibeam Echo Sounding Combined With Mass Conservation," *Geophysical Research Letters*, vol. 44, Nov. 2017.

[3] C. A. Greene, D. E. Gwyther, and D. D. Blankenship, "Antarctic Mapping Tools for Matlab," *Computers & Geosciences*, vol. 104, pp. 151–157, July 2017.

[4] D. J. Cavalieri, C. L. Parkinson, P. Gloersen, and H. J. Zwally, "Sea Ice Concentrations from Nimbus-7 SMMR and DMSP SSM/I-SSMIS Passive Microwave Data, Version 1," 1996, updated yearly.

[5] L. Padman, H. A. Fricker, R. Coleman, S. Howard, and L. Erofeeva, "A new tide model for the Antarctic ice shelves and seas," *Annals of Glaciology*, vol. 34, pp. 247–254, 2002.

[6] B. Smith, H. A. Fricker, A. S. Gardner, B. Medley, J. Nilsson, F. S. Paolo, N. Holschuh, S. Adusumilli, K. Brunt, B. Csatho, K. Harbeck, T. Markus, T. Neumann, M. R. Siegfried, and H. J. Zwally, "Pervasive ice sheet mass loss reflects competing ocean and atmosphere processes," *Science*, vol. 368, pp. 1239–1242, June 2020.

[7] H. De Angelis and P. Skvarca, "Glacier Surge After Ice Shelf Collapse," *Science*, vol. 299, pp. 1560–1562, Mar. 2003.

[8] M. Thoma, A. Jenkins, D. Holland, and S. Jacobs, "Modelling Circumpolar Deep Water intrusions on the Amundsen Sea continental shelf, Antarctica," *Geophysical Research Letters*, vol. 35, p. L18602, Sept. 2008.

[9] H. D. Pritchard, S. R. M. Ligtenberg, H. A. Fricker, D. G. Vaughan, M. R. van den Broeke, and L. Padman, "Antarctic ice-sheet loss driven by basal melting of ice shelves," *Nature*, vol. 484, pp. 502–505, Apr. 2012.

[10] F. S. Paolo, H. A. Fricker, and L. Padman, "Volume loss from Antarctic ice shelves is accelerating," *Science*, vol. 348, no. 6232, pp. 327–331, 2015.

[11] T. A. Scambos, "Glacier acceleration and thinning after ice shelf collapse in the Larsen B embayment, Antarctica," *Geophysical Research Letters*, vol. 31, no. 18, p. L18402, 2004.

[12] T. K. Dupont and R. B. Alley, "Assessment of the importance of ice-shelf buttressing to ice-sheet flow," *Geophysical Research Letters*, vol. 32, no. 4, 2005.

[13] E. Rignot, J. Mouginot, M. Morlighem, H. Seroussi, and B. Scheuchl, "Widespread, rapid grounding line retreat of Pine Island, Thwaites, Smith, and Kohler glaciers, West Antarctica, from 1992 to 2011," *Geophysical Research Letters*, vol. 41, pp. 3502–3509, May 2014.

[14] J. J. Fürst, G. Durand, F. Gillet-Chaulet, L. Tavard, M. Rankl, M. Braun, and O. Gagliardini, "The safety band of Antarctic ice shelves," *Nature Climate Change*, vol. 6, pp. 479–482, May 2016.

[15] R. C. Aster and J. P. Winberry, "Glacial seismology," *Reports on Progress in Physics*, vol. 80, p. 126801, Dec. 2017.

[16] Q. Kong, D. T. Trugman, Z. E. Ross, M. J. Bianco, B. J. Meade, and P. Gerstoft, "Machine Learning in Seismology: Turning Data into Insights," *Seismological Research Letters*, vol. 90, pp. 3–14, Jan. 2019.

[17] M. J. Bianco and P. Gerstoft, "Travel Time Tomography With Adaptive Dictionaries," *IEEE Transactions on Computational Imaging*, vol. 4, pp. 499–511, Dec. 2018.

[18] M. J. Bianco, P. Gerstoft, K. B. Olsen, and F.-C. Lin, "High-resolution seismic tomography of Long Beach, CA using machine learning," *Scientific Reports*, vol. 9, p. 14987, Dec. 2019.

[19] C. W. Johnson, H. Meng, F. Vernon, and Y. Ben-Zion, "Characteristics of Ground Motion Generated by Wind Interaction With Trees, Structures, and Other Surface Obstacles," *Journal of Geophysical Research: Solid Earth*, vol. 124, pp. 8519–8539, Aug. 2019.

[20] C. Bishop, *Pattern Recognition and Machine Learning*. Information Science and Statistics, Springer-Verlag New York, 1 ed., 2006.

[21] B. K. Holtzman, A. Paté, J. Paisley, F. Waldhauser, and D. Repetto, "Machine learning reveals cyclic changes in seismic source spectra in Geysers geothermal field," *Science Advances*, vol. 4, p. eaao2929, May 2018.

[22] C. W. Johnson, Y. Ben-Zion, H. Meng, and F. Vernon, "Identifying Different Classes of Seismic Noise Signals Using Unsupervised Learning," *Geophysical Research Letters*, vol. 47, Aug. 2020.

[23] M. G. Baker, R. C. Aster, R. E. Anthony, J. Chaput, D. A. Wiens, A. Nyblade, P. D. Bromirski, P. Gerstoft, and R. A. Stephen, "Seasonal and spatial variations in the ocean-coupled ambient wavefield of the Ross Ice Shelf," *Journal of Glaciology*, vol. 65, pp. 912–925, Dec. 2019.

[24] R. A. Bindschadler, M. A. King, R. B. Alley, S. Anandakrishnan, and L. Padman, "Tidally Controlled Stick-Slip Discharge of a West Antarctic Ice Stream," *Science*, vol. 301, pp. 1087–1089, Aug. 2003.

[25] R. A. Bindschadler, P. L. Vornberger, M. A. King, and L. Padman, "Tidally driven stick–slip motion in the mouth of Whillans Ice Stream, Antarctica," *Annals of Glaciology*, vol. 36, pp. 263–272, 2003.

[26] D. A. Wiens, S. Anandakrishnan, J. P. Winberry, and M. A. King, "Simultaneous teleseismic and geodetic observations of the stick–slip motion of an Antarctic ice stream," *Nature*, vol. 453, pp. 770–774, June 2008.

[27] C. G. Barcheck, S. Tulaczyk, S. Y. Schwartz, J. I. Walter, and J. P. Winberry, "Implications of basal micro-earthquakes and tremor for ice stream mechanics: Stick-slip basal sliding and till erosion," *Earth and Planetary Science Letters*, vol. 486, pp. 54–60, Mar. 2018.

[28] S. D. Olinger, B. P. Lipovsky, D. A. Wiens, R. C. Aster, P. D. Bromirski, Z. Chen, P. Gerstoft, A. A. Nyblade, and R. A. Stephen, "Tidal and Thermal Stresses Drive Seismicity Along a Major Ross Ice Shelf Rift," *Geophysical Research Letters*, vol. 46, pp. 6644–6652, June 2019.

[29] D. R. MacAyeal, A. F. Banwell, E. A. Okal, J. Lin, I. C. Willis, B. Goodsell, and G. J. MacDonald, "Diurnal seismicity cycle linked to subsurface melting on an ice shelf," *Annals of Glaciology*, vol. 60, pp. 137–157, Sept. 2019.

[30] J. Chaput, R. C. Aster, D. McGrath, M. Baker, R. E. Anthony, P. Gerstoft, P. Bromirski, A. Nyblade, R. A. Stephen, D. A. Wiens, S. B. Das, and L. A. Stevens, "Near-Surface Environmentally Forced Changes in the Ross Ice Shelf Observed With Ambient Seismic Noise," *Geophysical Research Letters*, vol. 45, Oct. 2018.

[31] P. D. Bromirski and R. A. Stephen, "Response of the Ross Ice Shelf, Antarctica, to ocean gravity-wave forcing," *Annals of Glaciology*, vol. 53, no. 60, pp. 163–172, 2012.

[32] P. D. Bromirski, Z. Chen, R. A. Stephen, P. Gerstoft, D. Arcas, A. Diez, R. C. Aster, D. A. Wiens, and A. Nyblade, "Tsunami and infragravity waves impacting A ntarctic ice shelves," *Journal of Geophysical Research: Oceans*, vol. 122, pp. 5786–5801, July 2017.

[33] Z. Chen, P. D. Bromirski, P. Gerstoft, R. A. Stephen, D. A. Wiens, R. C. Aster, and A. A. Nyblade, "Ocean-excited plate waves in the Ross and Pine Island Glacier ice shelves," *Journal of Glaciology*, vol. 64, pp. 730–744, Oct. 2018.

[34] M. G. Baker, R. C. Aster, D. A. Wiens, A. Nyblade, P. D. Bromirski, P. Gerstoft, and R. A. Stephen, "Teleseismic earthquake wavefields observed on the Ross Ice Shelf," *Journal of Glaciology*, pp. 1–17, Oct. 2020.

[35] Z. Chen, P. D. Bromirski, P. Gerstoft, R. A. Stephen, W. S. Lee, S. Yun, S. D. Olinger, R. C. Aster, D. A. Wiens, and A. A. Nyblade, "Ross Ice Shelf Icequakes Associated With Ocean Gravity Wave Activity," *Geophysical Research Letters*, vol. 46, pp. 8893–8902, Aug. 2019.

[36] A. Diez, P. Bromirski, P. Gerstoft, R. Stephen, R. Anthony, R. Aster, C. Cai, A. Nyblade, and D. Wiens, "Ice shelf structure derived from dispersion curve analysis of ambient

seismic noise, Ross Ice Shelf, Antarctica," *Geophysical Journal International*, vol. 205, pp. 785–795, May 2016.

[37] M. C. Hell, B. D. Cornelle, S. T. Gille, A. J. Miller, and P. D. Bromirski, "Identifying Ocean Swell Generation Events from Ross Ice Shelf Seismic Data," *Journal of Atmospheric and Oceanic Technology*, vol. 36, pp. 2171–2189, Nov. 2019.

[38] C. C. Aggarwal and C. K. Reddy, eds., *Data Clustering: Algorithms and Applications*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, Boca Raton: Chapman and Hall/CRC, 2014.

[39] S. M. Mousavi, S. P. Horton, C. A. Langston, and B. Samei, "Seismic features and automatic discrimination of deep and shallow induced-microearthquakes using neural network and logistic regression," *Geophysical Journal International*, vol. 207, pp. 29–46, Oct. 2016.

[40] D. T. Trugman and P. M. Shearer, "GrowClust: A Hierarchical Clustering Algorithm for Relative Earthquake Relocation, with Application to the Spanish Springs and Sheldon, Nevada, Earthquake Sequences," *Seismological Research Letters*, vol. 88, pp. 379–391, Mar. 2017.

[41] N. Riahi and P. Gerstoft, "Using graph clustering to locate sources within a dense sensor array," *Signal Processing*, vol. 132, pp. 110–120, Mar. 2017.

[42] L. Telesca and T. Chelidze, "Visibility Graph Analysis of Seismicity around Enguri High Arch Dam, Caucasus," *Bulletin of the Seismological Society of America*, vol. 108, pp. 3141–3147, Nov. 2018.

[43] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, (Berkeley, Calif.), pp. 281–297, University of California Press, 1967.

[44] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," *Applied Statistics*, vol. 28, no. 1, p. 100, 1979.

[45] M. Chamarczuk, Y. Nishitsuji, M. Malinowski, and D. Draganov, "Unsupervised Learning Used in Automatic Detection and Classification of Ambient-Noise Recordings from a Large-N Array," *Seismological Research Letters*, vol. 91, pp. 370–389, Jan. 2020.

[46] T. Perol, M. Gharbi, and M. Denolle, "Convolutional neural network for earthquake detection and location," *Science Advances*, vol. 4, p. e1700578, Feb. 2018.

[47] B. C. Wallet and R. Hardisty, "Unsupervised seismic facies using Gaussian mixture models," *Interpretation*, vol. 7, p. 19, Aug. 2019.

[48] L. Seydoux, R. Balestriero, P. Poli, M. de Hoop, M. Campillo, and R. Baraniuk, "Clustering earthquake signals and background noises in continuous seismic data with unsupervised deep learning," *Nature Communications*, vol. 11, p. 3972, Dec. 2020.

[49] S. J. Gibbons and F. Ringdal, "The detection of low magnitude seismic events using array-based waveform correlation," *Geophysical Journal International*, vol. 165, pp. 149–166, Apr. 2006.

[50] E. Beaucé, W. B. Frank, and A. Romanenko, "Fast Matched Filter (FMF): An Efficient Seismic Matched-Filter Search for Both CPU and GPU Architectures," *Seismological Research Letters*, vol. 89, pp. 165–172, Jan. 2018.

[51] C. J. Chamberlain, C. J. Hopp, C. M. Boese, E. Warren-Smith, D. Chambers, S. X. Chu, K. Michailos, and J. Townend, "EQcorrscan: Repeating and Near-Repeating Earthquake Detection and Analysis in Python," *Seismological Research Letters*, vol. 89, pp. 173–181, Jan. 2018.

[52] C. E. Yoon, O. O'Reilly, K. J. Bergen, and G. C. Beroza, "Earthquake detection through computationally efficient similarity search," *Science Advances*, vol. 1, p. e1501057, Dec. 2015.

[53] K. J. Bergen and G. C. Beroza, "Detecting earthquakes over a seismic network using single-station similarity measures," *Geophysical Journal International*, vol. 213, pp. 1984–1998, June 2018.

[54] A. J. Hotovec-Ellis and C. Jeffries, "Near Real-time Detection, Clustering, and Analysis of Repeating Earthquakes: Application to Mount St. Helens and Redoubt Volcanoes," Apr. 2016.

[55] H. M. Cole, "Tidally Induced Seismicity at the Grounded Margins of the Ross Ice Shelf, Antarctica," Master's thesis, Colorado State University, Fort Collins, Colorado, 2020.

[56] R. E. Bellman, *Adaptive Control Processes: A Guided Tour*. Rand Corporation, 1961.

[57] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning Series, Cambridge, MA: MIT Press, 2012.

[58] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the Surprising Behavior of Distance Metrics in High Dimensional Space," in *Database Theory — ICDT 2001* (G. Goos, J. Hartmanis, J. van Leeuwen, J. Van den Bussche, and V. Vianu, eds.), vol. 1973, pp. 420–434, Berlin, Heidelberg: Springer Berlin Heidelberg, 2001.

[59] M. Steinbach, L. Ertöz, and V. Kumar, "The Challenges of Clustering High Dimensional Data," in *New Directions in Statistical Physics: Econophysics, Bioinformatics, and Pattern Recognition* (L. T. Wille, ed.), pp. 273–309, Berlin, Heidelberg: Springer Berlin Heidelberg, 2004.

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

[61] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016.

[62] T. A. Reddy, K. R. Devi, and S. V. Gangashetty, "Nonlinear principal component analysis for seismic data compression," in *2012 1st International Conference on Recent Advances in Information Technology (RAIT)*, (Dhanbad, India), pp. 927–932, IEEE, Mar. 2012.

[63] G. E. Hinton, "Reducing the Dimensionality of Data with Neural Networks," *Science*, vol. 313, pp. 504–507, July 2006.

[64] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised Deep Embedding for Clustering Analysis," *Proceedings of the 33rd international conference on machine learning*, p. 10, 2016.

[65] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion," *Journal of Machine Learning Research*, p. 38, Dec. 2010.

[66] E. Min, X. Guo, Q. Liu, G. Zhang, J. Cui, and J. Long, "A Survey of Clustering With Deep Learning: From the Perspective of Network Architecture," *IEEE Access*, vol. 6, pp. 39501–39514, 2018.

[67] S. E. Chazan, S. Gannot, and J. Goldberger, "Deep Clustering Based on a Mixture of Autoencoders," *arXiv:1812.06535 [cs, stat]*, Mar. 2019.

[68] A. Boubekki, M. Kampffmeyer, U. Brefeld, and R. Jenssen, "Joint optimization of an autoencoder for clustering and embedding," *Machine Learning*, vol. 110, pp. 1901–1937, July 2021.

[69] S. M. Mousavi, W. Zhu, W. Ellsworth, and G. Beroza, "Unsupervised Clustering of Seismic Signals Using Deep Convolutional Autoencoders," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, pp. 1693–1697, Nov. 2019.

[70] D. Snover, C. W. Johnson, M. J. Bianco, and P. Gerstoft, "Deep Clustering to Identify Sources of Urban Seismic Noise in Long Beach, California," *Seismological Research Letters*, vol. 92, pp. 1011–1022, Mar. 2021.

[71] E. Ozanich, A. Thode, P. Gerstoft, L. A. Freeman, and S. Freeman, "Deep embedded clustering of coral reef bioacoustics," *J. Acoust. Soc. Am.*, p. 16, 2021.

[72] R. Allen, "Automatic phase pickers: Their present use and future prospects," *Bulletin of the Seismological Society of America*, vol. 72, pp. S225–S242, Dec. 1982.

[73] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient BackProp," in *Neural Networks: Tricks of the Trade: Second Edition* (G. Montavon, G. B. Orr, and K.-R. Müller, eds.), pp. 9–48, Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[74] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980 [cs]*, Jan. 2017.

[75] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, 2008.

[76] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *The Annals of Mathematical Statistics*, vol. 22, pp. 79–86, Mar. 1951.

[77] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, pp. 411–423, May 2001.

[78] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, Nov. 1987.

[79] J. P. Nicolas, A. M. Vogelmann, R. C. Scott, A. B. Wilson, M. P. Cadeddu, D. H. Bromwich, J. Verlinde, D. Lubin, L. M. Russell, C. Jenkinson, H. H. Powers, M. Ryczek, G. Stone, and J. D. Wille, "January 2016 extensive summer melt in West Antarctica favoured by strong El Niño," *Nature Communications*, vol. 8, p. 15799, Aug. 2017.

[80] E. Klein, C. Mosbeux, P. D. Bromirski, L. Padman, Y. Bock, S. R. Springer, and H. A. Fricker, "Annual cycle in flow of Ross Ice Shelf, Antarctica: Contribution of variable basal melting," *Journal of Glaciology*, vol. 66, pp. 861–875, Oct. 2020.

[81] D. Wiens and P. Bromirski, "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," 2014.

[82] M. Beyreuther, R. Barsch, L. Krischer, T. Megies, Y. Behr, and J. Wassermann, "ObsPy: A Python Toolbox for Seismology," *Seismological Research Letters*, vol. 81, pp. 530–533, May 2010.

Chapter 3

Analysis of underwater acoustic data collected under sea ice during the Useful Arctic Knowledge 2021 cruise

Acoustical observations are presented from the Useful Arctic Knowledge (UAK) 2021 cruise, an early-career training program which took place in sea ice north of Fram Strait in June 2021 on board the Norwegian Coast Guard icebreaker KV *Svalbard*. Through oceanographic sampling and three acoustics-related tasks, participants were introduced to practical applications of underwater acoustics and observed the ocean environment and its role in acoustic propagation. Propagation modeling from oceanographic sampling confirmed an upward-refracting environment throughout the cruise. In the first task, a drifting acoustic receiver buoy with a single hydrophone was deployed in an ice floe and left to passively record for eight days. Various biological sounds were recorded, including bearded seals and cetaceans. In the second task, acoustic localization using an active pinger system was used to recover an operational oceanographic mooring from the seabed. The third task involved passive acoustic observations beneath sea ice whenever KV *Svalbard* fastened herself to ice floes and measurements of the ice were made. Through the UAK 2021 cruise, participants learned the utility and value of using underwater acoustics for operations in the Arctic Ocean, particularly in location and retrieval of equipment and in measuring and sensing the environment.

3.1 Introduction

The Useful Arctic Knowledge (UAK) program is an international, interdisciplinary program hosted by Norway's Nansen Environmental and Remote Sensing Center (NERSC), intended to build and maintain strong partnerships among students, early career scientists, and experienced experts in selected Arctic topics.One of the defining features of the program is an annual scientific cruise. In June 2021, participants from nine countries embarked on board the Norwegian Coast Guard icebreaker KV *Svalbard* and conducted observations of sea ice, acoustic measurements, conductivity-salinity-temperature (CTD) casts, mooring recovery in sea ice, buoy deployments, remote sensing product analysis, and sea ice navigation. This paper presents observations and discussion from tasks related to acoustic observation.

Acoustic and oceanographic observations were made over the course of eight days utilizing CTD casts and passive and active acoustics. On 8 June, a drifting acoustic receiver buoy with several instruments was deployed in an ice floe and left to passively record for eight days. An acoustic localization of the buoy was demonstrated before the *Svalbard* continued on with her cruise. At various points in the subsequent week, *Svalbard* fastened herself to ice floes where measurements of the ice were made. At each of these ice stations, passive acoustic observations were made. On 12 June, acoustic localization was used to successfully estimate the position of an oceanographic mooring for recovery. On 16 June, the *Svalbard* located and recovered the drifting acoustic receiver buoy.

The purpose of the acoustic tasks was to introduce participants to practical applications of underwater acoustics, including equipment selection and preparation, mooring construction, deck handling, and collection and handling of data. Furthermore, participants observed the ocean environment and its role in underwater acoustic propagation. Finally, participants learned the utility and value of using underwater acoustics for operations in the Arctic Ocean, particularly in location and retrieval of equipment and in measuring the environment. In the sections that follow, a description of the acoustic environment is provided, and each acoustic task is discussed

in detail.

3.2 Environment

3.2.1 Environmental sampling

The Arctic Ocean is generally an upward refracting environment [1]. Unlike more temperate latitudes where water at the surface is warmer than at depth, surface waters in the Arctic are typically cooler and fresher, resulting in lower sound speeds, although intrusion of warmer, more saline water does occur at inflows in the Fram Strait and Bering Strait [2, 3]. In Fram Strait, a great deal of oceanographic variability occurs due to Atlantic Water transported by the West Spitsbergen Current [2], which has been measured directly with oceanographic measurements [4] and acoustically in various tomography experiments [5, 6]. In the Arctic, the ambient noise field is related to transients generated by sea ice as well as biological activity and anthropogenic sources such as seismic airgun surveys [7].

As part of an ongoing series of oceanographic measurements in Fram Strait and north of Svalbard, eight CTD casts were performed at various points throughout the cruise. Figure 3.1a shows sampling locations and results for each CTD cast. Casts 1 through 7 show typical sound speed profiles (SSP) for the Arctic Ocean, with the top 100 m strongly influenced by relatively cold, fresh water. In cast 8, the strength of the layer is substantially diminished, likely as a result of Atlantic Water intrusion through Fram Strait.

Additional CTD data were collected using expendable bathythermographs (XBT) and were consistent with profiles shown in Fig. 3.1a, and a continuously recording CTD instrument was mounted to the drifting acoustic receiver buoy (Sec. 3.3).

3.2.2 Acoustic propagation modeling

CTD profiles revealed an upward refracting environment which led to a surface duct in the upper sonic layer. A combination of relatively fresh water combined with water cooled

by the Arctic air results in especially slow sound speeds in the upper ocean. Figure 3.1b shows a simplified acoustic propagation model for transmission loss (TL) using environmental data collected by an XBT during an acoustic localization exercise (Sec. 3.4). The KRAKEN normal mode propagation model [8] was used to compute transmission loss for a 150 Hz source positioned at a depth of 30 m, which is the approximate depth of the hydrophone on the receiver buoy (Sec. 3.3). TL is shown out to a range of 20 km (Fig. 3.1b, upper panel) and 100 km (Fig. 3.1b, lower panel). Surface ducting is visible in both panels, with the sonic layer depth located at the thermocline at a depth of approximately 150 m. The SSP below the thermocline was also positive, and in the lower panel of Fig. 3.1b there is a half channel with annuli occuring approximately every 35 km. A 2 m layer of ice is assumed, but this modeling does not take into account ice roughness, thus TL is underestimated as ice scatters and reflects sound into the ocean bottom. Nevertheless, the model shows that the dominant propagation paths are direct path (close range), surface duct (medium range due to the frequent interactions with the ice and surface scattering), and half-channel surface bounce, especially for smaller launch angles in deeper water.

3.3 Drifting Acoustic Receiver Buoy

3.3.1 Equipment, deployment, and recovery

The drifting acoustic receiver buoy consisted of a weighted line approximately 35 m long suspended from a float. A Multi-électronique μAURAL recorder with an integrated HTI 96-min hydrophone, sampling continuously at 48 kHz sampling frequency, was mounted on the line at 30 m depth. A Sea-Bird Scientific SBE37 CTD, sampling with an interval of five minutes, was mounted at 33 m depth. A XEOS GPS receiver was fastened to the float itself, and recorded the position of the float every hour. An Edgetech Coastal Acoustic Transponder (CAT) was mounted for acoustic localization.

The buoy was deployed on 8 June 2021 at 17:52 CET in an ice floe located at 80°57.855′N

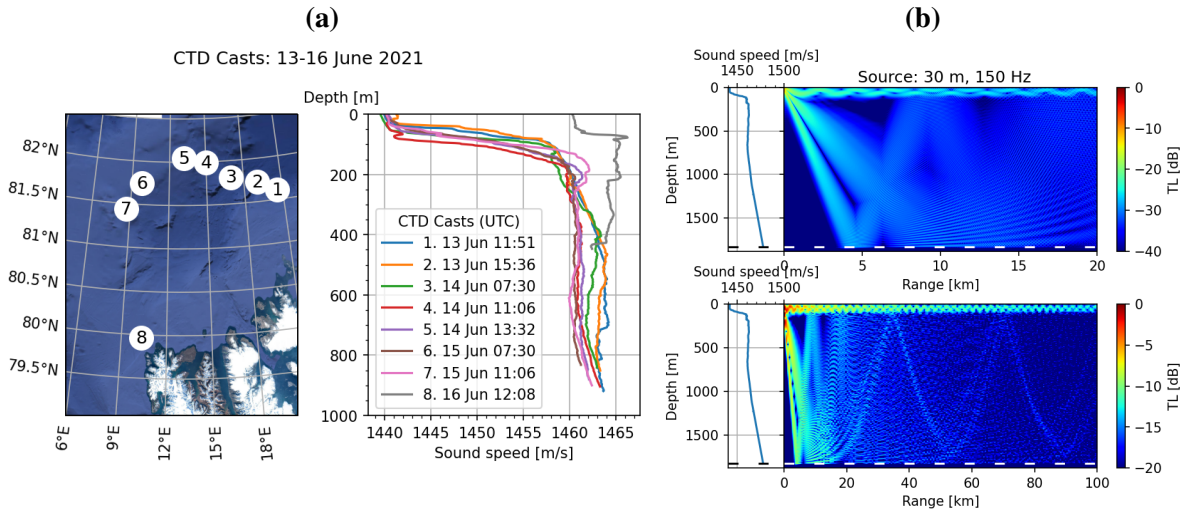

Figure 3.1. (a) Conductivity/temperature/depth (CTD) instrument casts taken throughout the cruise. (b) Transmission loss for a source at 10.5 m, 200 Hz using KRAKEN normal mode propagation model for 20 km (top) and 100 km (bottom). The sound speed profile used for the model is shown in the left panels and is from an XBT shot during the acoustic localization training described in Sec. 3.4.

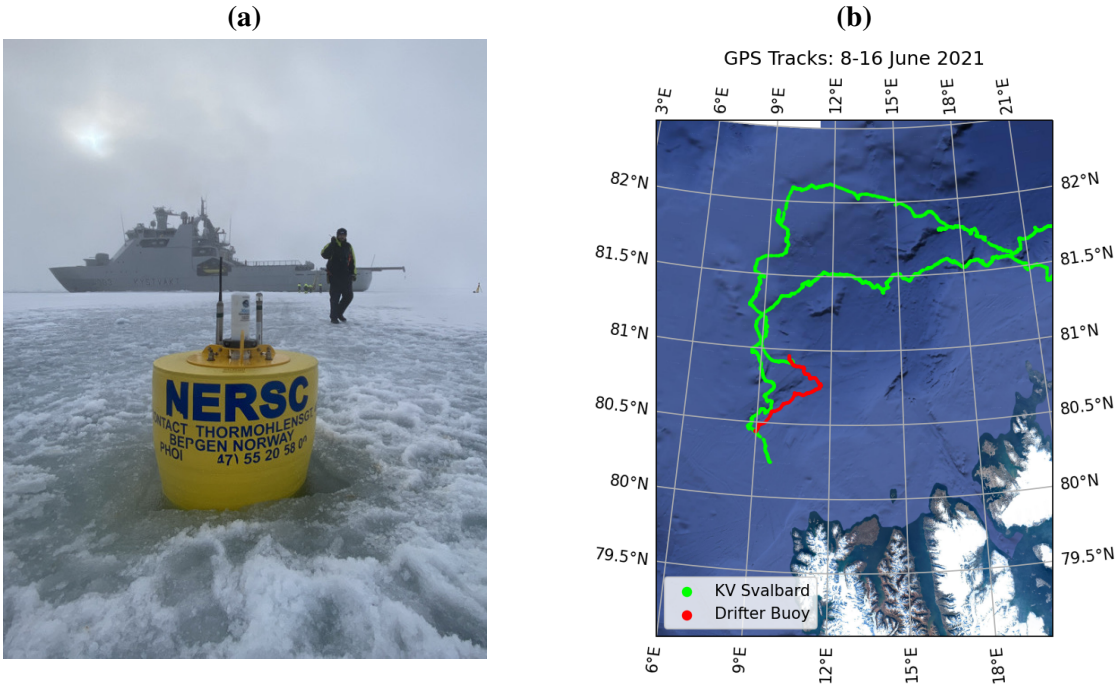

Figure 3.2. (a) Deployment of the drifting acoustic receiver buoy on an ice floe. (Photo: William Jenkins) (b) The track of KV *Svalbard* (green) and the buoy (red) are shown for the duration of the buoy deployment.

and 010°09.437′E. To ensure the buoy remained fixed to the ice floe, a hole was drilled through the ice large enough for the instruments to pass through, but small enough that the float would remain lodged at the surface. In the event the ice floe were to melt or fail, the float ensured the instruments would not sink. The buoy deployment is depicted in Fig. 3.2a.

The drifting buoy and hydrophone were retrieved from the ice floe on 16 June 2021 at 09:09 CET at 80°26.045′N and 008°59.338′E. The buoy's GPS transceiver provided general localization, and as the *Svalbard* approached, the buoy was identified visually. The buoy had cumulatively traveled 108.7 km, and as seen in Fig. 3.2b, initially drifted southeast, then changed directions to the southwest.

3.3.2 Data analysis

The ice floe that the drifting acoustic receiver buoy was fastened to thinned over the course of the deployment, with a measured ice thickness of about 70 cm at the start of the deployment and about 30 cm upon recovery. Previous studies suggest melting sea ice produces sound underwater in the frequency range of a few hundred Hz to a few kHz [9]. To assess whether this signal is observable in our data, we compare the sound power measured by the hydrophone in several frequency bands to measured or modeled (ERA5 reanalysis [10]) environmental variables which could reasonably be expected to be indicative of melting of the sea ice near the buoy.

It is apparent visually from Fig. 3.3, and from the correlations presented in Table 3.1, that the dominant feature in the integrated sound power is an increase in the energy present at low frequencies corresponding to times when the over-ground speed of the buoy and the ERA5 10 m wind speed. By listening to the recordings, it was determined that most of this energy is likely a result of strum noise due to the motion of the buoy through the water. The depth of the CTD, computed from the measured pressure, can be seen to decrease at the periods with the highest speed over ground, suggesting that there was sufficient force being exerted on the cable and instrumentation to cause the line to tilt away from vertical, which adds support for the hypothesis of strum noise. It seems likely that the correlation between wind speed and sound

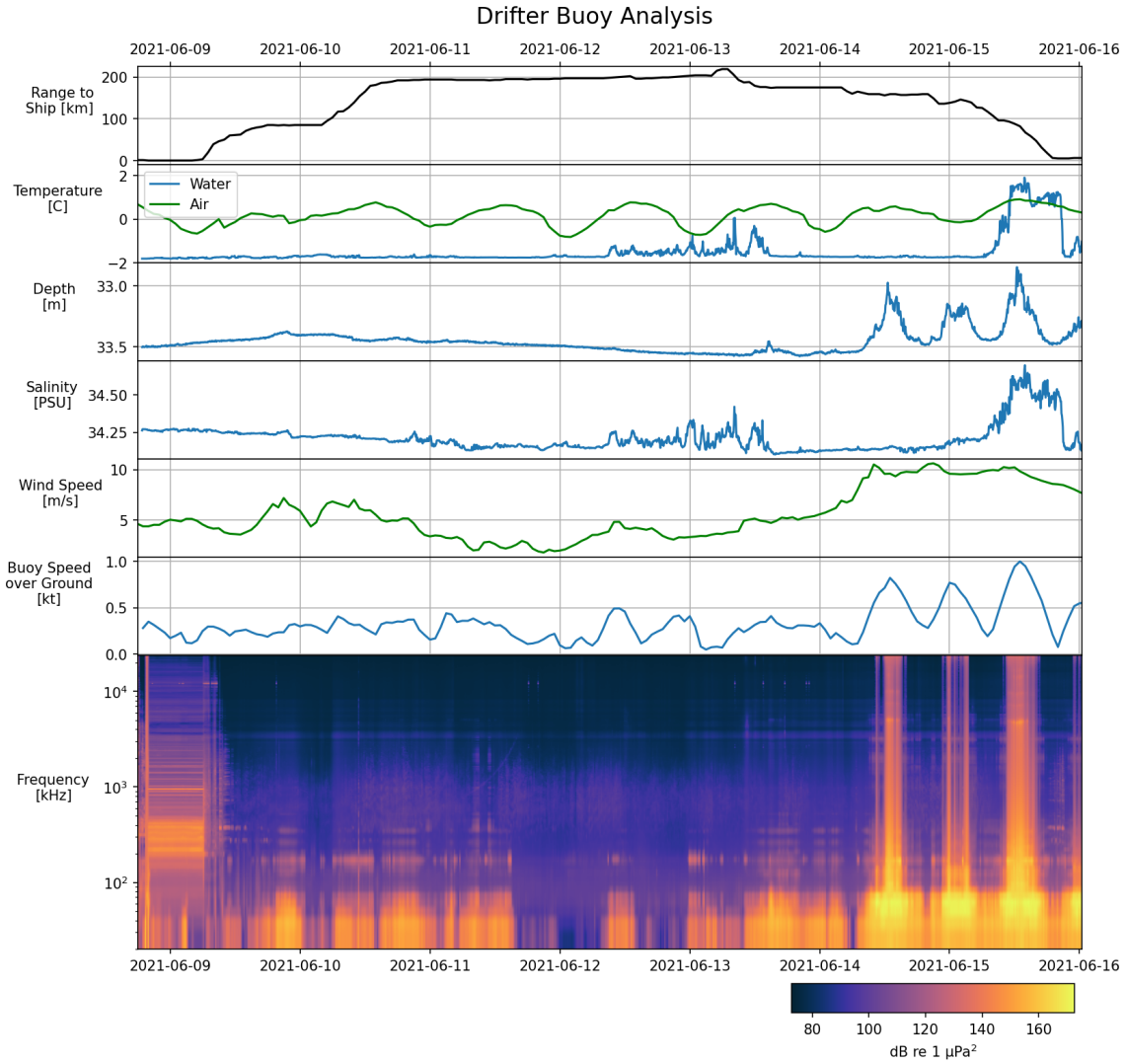

Figure 3.3. Time series of various environmental variables, which were measured on the drifting acoustic receiver buoy by the CTD (water temperature, salinity), GPS (buoy speed over ground, range to ship), or taken from ERA5 reanalysis products (air temperature, wind speed). The bottom panel is a spectrogram of the acoustic data recorded by the hydrophone over the duration of the deployment, showing the distribution of energy over frequencies. Note that the sharp vertically uniform bands of high energy are periods where the sound was sufficiently loud to saturate the recording; these periods have accordingly been removed from the analysis presented in Table 3.1.

power at low frequencies is at least partially a result of the causal relationship between wind speed and the speed of the ice flows, which determines the speed of the buoy. However, the weak but still elevated correlations between wind speed and sound power even at high frequencies,

Table 3.1. Correlation between various environmental factors and the logarithm of sound power integrated over different frequency bands.

	Frequency (kHz)				
Parameter	0-0.3	0.3-1	1-3	3-10	10-24
Water temperature	0.1004	0.0662	0.0943	0.1351	-0.0093
Buoy speed over ground	0.4707	-0.0241	0.0603	0.0828	-0.0950
ERA5 10 m wind speed	0.6358	0.1232	0.2302	0.3216	0.0169
ERA5 2 m air temperature	0.2349	-0.0900	-0.0452	-0.0272	-0.1085

where buoy speed over ground and sound power are completely uncorrelated, suggests that wind acting on the available area of open water could be causing some increase in underwater noise.

Even outside the low-frequency band dominated by the strum noise, we find no clear relationship between the environmental variables that might be indicative of melting and the sound power in any frequency band. The slight rise in water temperature during the middle of the record does not correspond to any discernible increase in sound power, and the rise in water temperature near the end of the record unfortunately corresponds with a period of hydrophone saturation.

3.3.3 Highlights

In this section, we highlight several signals recorded by the drifting acoustic receiver buoy that were representative of the broader soundscape observed during the recording period. The biologic examples that follow were excerpted from the morning of 12 June, when the record had the lowest ambient and self noise. During this quiet period, many types of marine mammals and fish vocalizations could be heard. The Discovery of Sound in the Sea website [11] was used to identify animals based on vocalization characteristics; however, the audio examples on the website are somewhat limited, and it was not possible to positively identify all of the vocalizations.

Though the data from the morning of 12 June can be readily analyzed for biologic sources, it is noteworthy that, in spite of the noise of the strumming buoy wire, biologic sources

were detected throughout the entirety of the record. The most prominent and easily identifiable sources were bearded seals. Cetaceans were also heard throughout the record and were most likely narwhals or beluga whales. Throughout the record, broadband pops and clicks were observed, some of which were likely associated with echolocation from marine mammals, although species attribution based on a single impulse was not possible. In addition to biologic sources, anthropogenic sounds such as machinery, propulsion, and icebreaking from the *Svalbard* are heard in the first two days of the record.

KV *Svalbard* self-noise

After deploying the buoy, KV *Svalbard* continued to drift with the ice floe overnight with her propulsion secured. Of course, various hotel loads and machinery continued operating. In recordings made near the *Svalbard*, including those from later ice stations, a periodic clicking noise was heard and is shown in the spectrogram in Fig. 3.4a, with a stronger click followed by a weaker click. The period of the full cycle was 1.75 s. Owing to the regular periodicity of this signal, it is likely generated by a rotating piece of machinery operating at approximately 34.3 RPM.

KV *Svalbard* was underway again on the morning of 9 June to conduct acoustic ranging to the buoy. Fig. 3.4b shows a spectrogram of the *Svalbard* maneuvering through sea ice. Below 1.5 kHz the propulsion machinery dominates the spectrum. Above 1.5 kHz, the spectrum becomes saturated with broadband noise as the ship impinges on ice and breaks it apart. Periods when the ship was in open water are indicated by regions of low energy above 1.5 kHz.

Bearded seals

Bearded seals were heard constantly throughout the entire record. These seals emit a warbling sound that sweeps downward in frequency, with occasional upshifts. From Fig. 3.4c, the duration of some of these calls was greater than one minute. As June was in the middle of their mating season, these calls may be attempts to attract mates. Bearded seals were spotted

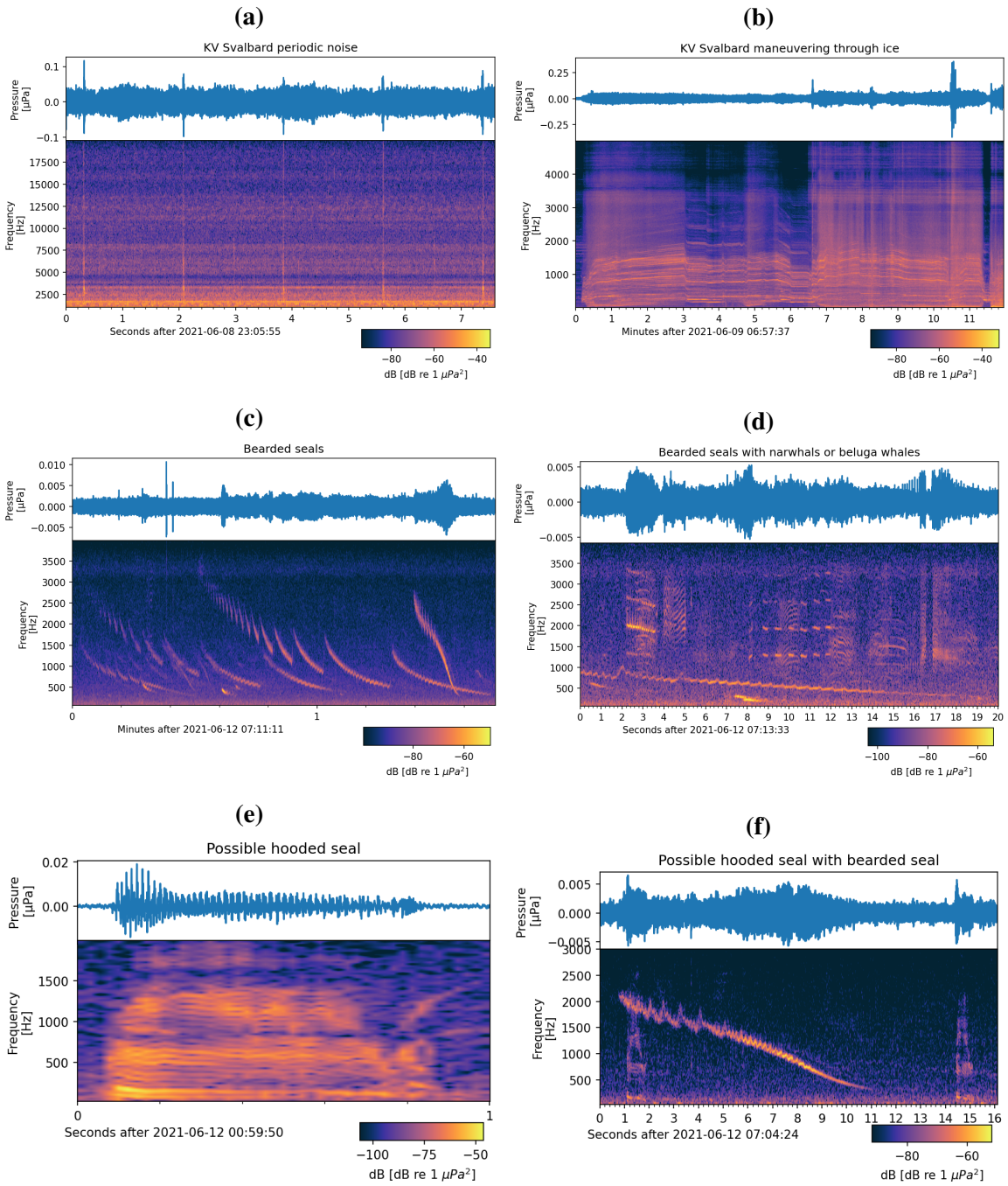

Figure 3.4. (a) Periodic transients from KV *Svalbard* suggest rotating machinery as the source. (b) KV *Svalbard* maneuvering through sea ice. (c) Bearded seal vocalizations. (d) Marine mammal vocalizations, including a downsweep made by a bearded seal (below 1 kHz) and vocalizations from narwhals or beluga whales. (e) Possible hooded seal vocalization. (f) Possible hooded seal vocalizations appear at the beginning and end of the spectrogram, shown with an intervening bearded seal call.

from the *Svalbard* on sea ice throughout the cruise.

Narwhals or beluga whales

Figure 3.4d shows examples of vocalizations from several biologic sources. The long, warbling downsweep of a bearded seal is visible below 1 kHz over the course of the entire spectrogram. Fine striations above 1 kHz sweeping rapidly up and down are likely narwhals or beluga whales, and were heard throughout the entire record. A vocalization with fundamental frequency at approximately 1.3 kHz and harmonics at 700 Hz intervals is observed with a single call at 2 s, followed by a repetitive train of pulses between 8 and 12 s; this, too, was likely a a narwhal or beluga.

Possible hooded seals

One of the biologic sources was recorded on numerous occasions between 10-13 June appears to be a call from a hooded seal. These animals would emit single vocalizations as well as call repeatedly two to three times per minute for several minutes. The vocalization sounds like a deep, nasal "wow," with the beginning of the signal shifting downward in frequency, and at the very end sweeping up. Figure 3.4e shows the vocalization in detail, and Fig. 3.4f shows its co-occurrence with a bearded seal vocalization. While hooded seals are typically asocial, June marks the end of their mating season and these calls may be indicative of animals seeking a mate [12].

3.4 Acoustic Localization of Moorings

One of the primary acoustic tasks was to localize an oceanographic mooring for recovery. This is an especially challenging activity in the Arctic that requires careful coordination between the ship's bridge, deck hands, and acoustic operators. First, a patch of sea ice must be cleared over the mooring. This can be done by icebreaking or waiting for a lead to pass over the mooring. Second, the ship must triangulate the position of the mooring using acoustic two-way travel

times. Finally, once the mooring is localized, an acoustic release activates and the mooring floats to the surface for recovery. The entire procedure must occur as swiftly as possible since the sea ice is moving with the wind and current. Thus, the validity of the localization quickly becomes invalid, and the mooring could come up beneath the sea ice when released.

3.4.1 Equipment

An Edgetech CAT pinger worked in tandem with a shipboard transducer to perform active acoustic localization. On board KV *Svalbard*, a deck unit drove a transducer that produced an 11 kHz outbound signal. The CAT, upon receiving the 11 kHz signal, transmitted a 12 kHz signal. On board *Svalbard*, the time elapsed between the transmission of the 11 kHz signal and receipt of the 12 kHz signal constituted the two-way travel time. Assuming an average speed of sound in water, the distance between the source and receiver could be estimated. In Fig. 3.5a, a spectrogram of the sequence of localization signals is shown.

3.4.2 Methodology

Two-way travel times Δt were measured as the time T elapsed between the transmission of the 11 kHz signal and reception of the 12 kHz signal, plus a correction T^* to account for the

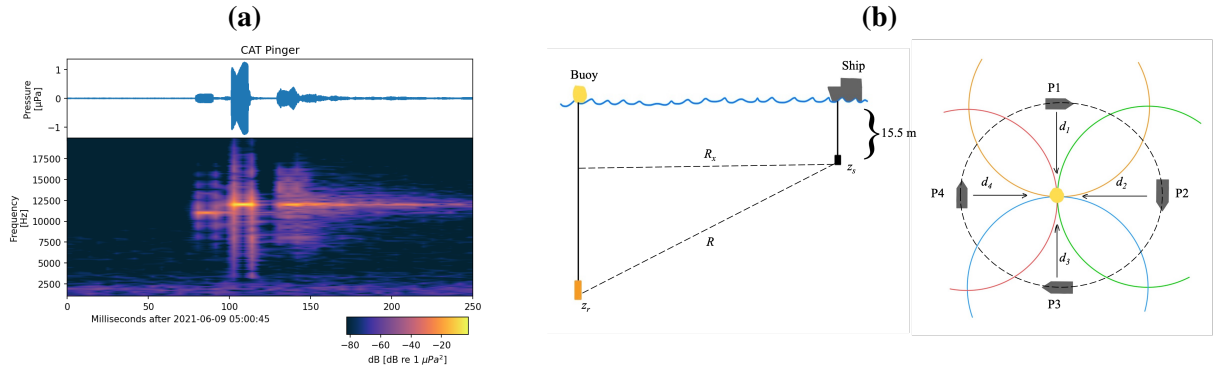

Figure 3.5. (a) Acoustic localization signals. The 11 kHz tone was transmitted by KV *Svalbard*, and the 12 kHz tone was the response transmitted by the transducer on the drifting acoustic receiver buoy. (b) Active acoustic localization betwen a ship and a buoy. R_x is used to plot range rings around the ship's position, and the buoy is localized at the intersection of multiple range rings.

delay on board the CAT pinger betwen its reception of the 11 kHz signal and transmission of the 12 kHz signal. The one-way travel time is then:

$$\Delta t = \frac{T}{2} - T^*. \tag{3.1}$$

Slant range R between the ship and the CAT pinger is then:

$$R = c_{\text{avg}} \Delta t, \tag{3.2}$$

where c_{avg} is the average sound speed between the two. Finally, the horizontal range R_x is given by:

$$R_x = \sqrt{R^2 - |z_r - z_s|^2}, \tag{3.3}$$

where z_r and z_s are the depths of the CAT pinger and ship's transducer, respectively.

Since the ship's transducer is a single, omnidirectional element, it is impossible to resolve bearing to the CAT pinger with a single measurement. To localize the buoy, several measurements are taken at different positions, with circles of radius R_x plotted at each position of measurement. Where the circles intersect indicates the estimated position of the mooring [13]. A schematic of the localization procedure is shown in Fig. 3.5b.

3.4.3 Results

Practice localization on known position

Acoustic localization was tested on the drifting acoustic receiver buoy on 9 June 2021. With a GPS receiver mounted on top of the buoy, the localization results could be compared to the actual positions of the buoy and ship. Two localizations were performed at ranges of approximately 500 m and 1000 m. For each localization, four measurements were taken at approximately 90° intervals around the buoy.

At each measurement position, the shipboard acoustic transducer was lowered to $z_s =$

Table 3.2. Example of two-way travel time data collected during acoustic localization.

	Ship Position			
Time	Latitude	Longitude	T (ms)	R_x (m)
05:02	80°55.3501'N	010°13.9118'E	851	627
05:15	80°55.1844'N	010°16.5782'E	656	478
05:35	80°55.4784'N	010°18.3356'E	865	636
05:53	80°55.5761'N	010°16.4029'E	795	589

15.5 m below the surface of the water. A piece of tape was attached to the wire marking how far the source needed to be lowered so that the depth would be consistent between measurements. After the transponder was lowered to the correct depth, the position of the ship was recorded and the 11 kHz signal was transmitted. Once the 12 kHz signal from the CAT arrived at the ship, the two-way travel time T was recorded and a time delay $T^* = 12.5$ ms used to obtain Δt. From XBT casts performed in the area, average SSP was $c_{avg} = 1442$ m/s. Using the CAT pinger depth of $z_r = 33$ m, horizontal range was calculated at each measurement position using Eq. 3.1–3.3. Table 3.2 includes an example of data collected during the localization.

At this point, participants were divided into two groups to estimate the position of the buoy for the 500 m localization and 1000 m localization. Using the horizontal ranges, each group plotted its range circles, but exact intersections were not obtained due to measurement errors and other factors. The most significant of these was that, because of the wind and currents, the ship and buoy were moving at slightly different velocities which were not accounted for in the calculations. Despite these errors, the experiment still produced areas of intersection where the buoy had the highest probability of being located. The two groups then independently developed ways to account for these errors and the effect of time elapsed between measurements, incorporating a linear rate of ice drift to estimate the buoy's position. KV *Svalbard*'s track, both buoy localization estimates, and the hourly GPS positions reported by the buoy are shown in Fig. 3.6. For the 500 m localization, there was an estimated error of 70.3 m, while for the 1000 m localization there was an estimated error of 81.5 m. Since during recovery operations cleared sea

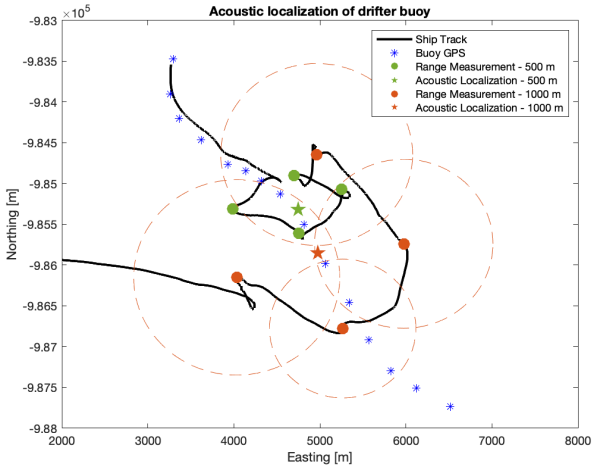

Figure 3.6. Results are shown from acoustic localizations conducted at a range of 500 m and 1000 m. Drifter buoy GPS positions are hourly.

Figure 3.7. Oceanographic mooring CNRS23, shown here being recovered by KV *Svalbard*, was localized using active acoustics before being released from its anchor. (Photo: Sofia Vakhutinsky)

ice typically spans several hundred meters, these errors were within the tolerance for successfully recovering a mooring.

Recovery of an oceanographic mooring

On the afternoon of 11 June 2021, KV *Svalbard* took station over oceanographic mooring CNRS23, which was deployed in 2019 to observe Atlantic Water inflow into the Arctic Ocean [14]. Recovery was timed to coincide with the occurrence of a large lead of open water over the mooring. While the position of the mooring was known from its 2019 deployment, acoustic localization was performed to confirm its location prior to recovery. Once the location was confirmed, an acoustic release was activated and the mooring floated to the surface. A small boat was deployed to retrieve the mooring, and the mooring with its numerous oceanographic instruments was recovered with the ship's crane as shown in Fig. 3.7.

3.5 Ice Stations

Additional acoustic measurements were taken with a hydrophone lowered by hand through a hole drilled in the sea ice during sea ice station measurements. Most of the time, mechanical noise from the nearby ship was far louder than any environmental signals that might have been of interest. One exception is the controlled explosion, described here.

On the evening of 14 June, KV *Svalbard* stationed herself on an ice floe located at 82°05.607′N and 010°02.179′E. Though the main purpose of this visit was to observe ice ridges on the floe, an interesting opportunity for acoustic measurement arose when the ship's crew was given permission to detonate expired explosives. Using one of the many holes drilled for ice ridge observation, a hydrophone was lowered beneath the ice to capture the explosion as heard underwater. The explosives were emplaced in the ice approximately two hundred meters away.

Explosions create an impulsive signal, enabling an estimate of the bottom depth using acoustic travel times. The distance traveled by an acoustic wave is $d = ct$, where c is the speed of sound and t is the time elapsed. In this case, since the acoustic signal is traveling to the bottom of the ocean and back, t is the two-way travel time and must be divided by two to give the bottom depth z_b:

$$z_b = \frac{ct}{2} \tag{3.4}$$

Because the depth of the ocean is much greater than the distance between the source and receiver at the surface, the source and receiver are assumed to be in the same position. Figure 3.8 shows the normalized acoustic pressure and spectrogram of the recorded signal. The two-way travel time is obtained from the time difference of arrival between the first and second impulses. Using an estimated average sound speed of 1490 m/s and a measured time difference of arrival of 1.293 s, the bottom depth is estimated to be 963 m at this location, which is consistent with bathymetric data from the International Hydrographic Organization.

Figure 3.8 contains some noteworthy propagation features. The first wave packet appears

to contain multiple sub-packets of energy. This is likely a combination of two factors. First, several explosives were detonated, but due to latency in the detonation cord and fuses, they did not detonate simultaneously. The asynchronous detonation was recorded on film and audio by the observers. Second, because the explosives were emplaced within the ice, energy propagates seismo-acoustically through elastic media (the ice) as well as through the water. The speed of propagation for longitudinal waves in ice is approximately 3,800 m/s, and shear waves are approximately 1,800 m/s [15]. As these waves propagate outward through the ice, energy along this wavefront is transmitted into the water. These elastic modes of propagation are likely the first arrivals recorded, followed by direct path propagation through water between the source and the receiver. Since the ocean acts as an acoustic waveguide, the second packet of energy exhibits geometric dispersion, with an interference pattern taking shape as bands of energy through time and frequency following the arrival of the impulse. A third, weaker arrival remains impulsive, but is further attenuated and dispersed than the second arrival.

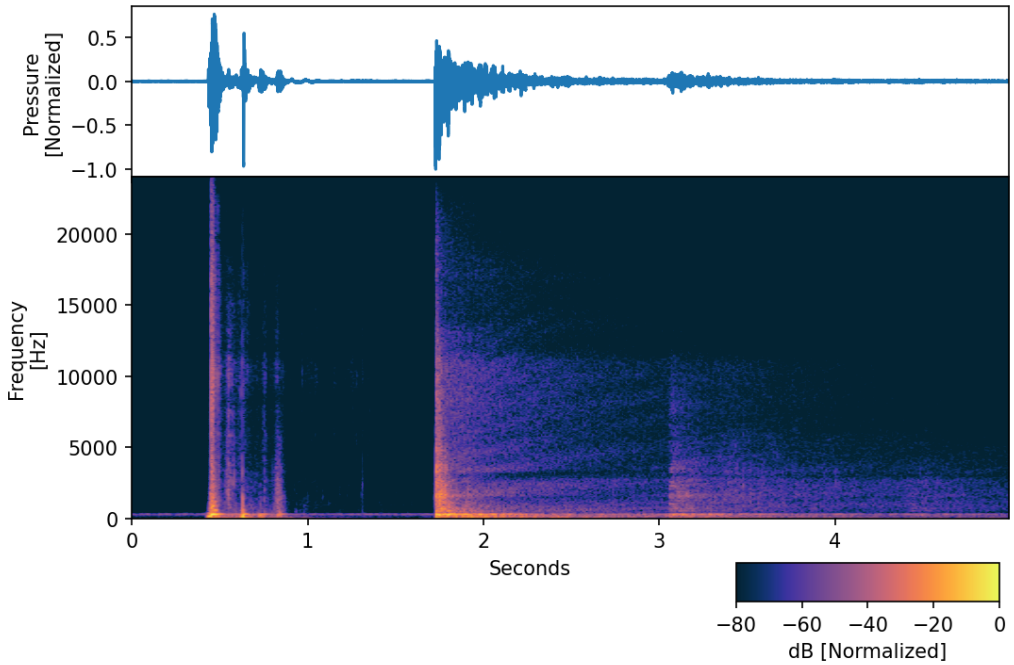

Figure 3.8. Normalized time series and spectrogram of explosives detonation on sea ice, followed by first and subsequent bottom bounce reflections.

3.6 Conclusion

The 2021 UAK cruise successfully demonstrated the use of both active and passive underwater acoustics for observation and localization in a challenging and dynamic sea ice environment. Additionally, oceanographic and atmospheric observations were used to model the acoustic propagation environment and to explain observations. Participants received hands-on experience at nearly every stage of the various acoustic tasks, and gained solid insights into the utility of underwater acoustics as both a practical tool for handling oceanographic equipment, as well as a method for observing the environment.

3.7 Acknowledgements

Mapping data provided by Google Earth. This work was made possible by: Integrated Arctic Observing System (INTAROS) Project H2020 (EU CORDIS grant no. 727890); Research Council of Norway (contract no. 274891); and Office of Naval Research (ONR) Global. The authors would like to thank the officers and crew of KV *Svalbard* for their highly professional support of this cruise. Participants were additionally instructed by Agnieszka Beszczynska-Möller (Institute of Oceanology, Polish Academy of Science), Alistair Everett (Norwegian Meteorological Institute), Tom Rune Lauknes (Norwegian Research Centre AS), and Elena McCarthy (ONR Global). Additional participants included Astrid Stallemo (NERSC); Anna Mathea Skaar, Mads Skjerven Moldrheim, Güney Dinçtürk, and Elinor Tessin (University of Bergen); and Anna Telegina, Laust Færch, and Jozef Rusin (University of Tromsø).

Chapter 3, in full, is a reprint of the material as it appears in the Proceedings of Meetings on Acoustics 2022. Jenkins, William; Johnson, Hayden; Vakhutinsky, Sofia; Helmberger, Meghan; Storheim, Espen; Sagen, Hanne; Sandven, Stein, Acoustical Society of America, 2022. The dissertation author was the primary investigator and author of this paper.

3.8 References

[1] R. J. Urick, *Principles of Underwater Sound*. Peninsula Publishing, 1983.

[2] L. D. Talley, G. L. Pickard, W. J. Emery, and J. H. Swift, *Descriptive Physical Oceanography: An Introduction*. Academic Press, 2011.

[3] P. F. Worcester and M. S. Ballard, "Ocean acoustics in the changing Arctic," *Physics Today*, vol. 73, pp. 44–49, Dec. 2020.

[4] M. D. Pérez-Hernández, R. S. Pickart, D. J. Torres, F. Bahr, A. Sundfjord, R. Ingvaldsen, A. H. H. Renner, A. Beszczynska-Möller, W.-J. Appen, and V. Pavlov, "Structure, Transport, and Seasonality of the Atlantic Water Boundary Current North of Svalbard: Results From a Yearlong Mooring Array," *Journal of Geophysical Research: Oceans*, vol. 124, pp. 1679–1698, Mar. 2019.

[5] H. Sagen, B. D. Dushaw, E. K. Skarsoulis, D. Dumont, M. A. Dzieciuch, and A. Beszczynska-Möller, "Time series of temperature in Fram Strait determined from the 2008–2009 DAMOCLES acoustic tomography measurements and an ocean model," *Journal of Geophysical Research: Oceans*, vol. 121, pp. 4601–4617, July 2016.

[6] H. Sagen, P. F. Worcester, M. A. Dzieciuch, F. Geyer, S. Sandven, M. Babiker, A. Beszczynska-Möller, B. D. Dushaw, and B. Cornuelle, "Resolution, identification, and stability of broadband acoustic arrivals in Fram Strait," *The Journal of the Acoustical Society of America*, vol. 141, pp. 2055–2068, Mar. 2017.

[7] E. Ozanich, P. Gerstoft, P. F. Worcester, M. A. Dzieciuch, and A. Thode, "Eastern Arctic ambient noise on a drifting vertical array," *The Journal of the Acoustical Society of America*, vol. 142, pp. 1997–2006, Oct. 2017.

[8] M. B. Porter, "The KRAKEN normal mode program," *SACLANT Undersea Research Centre Memorandum (SM-245)/Naval Research Laboratory Memorandum Report 6920*, Sept. 1991.

[9] M. M. Mahanty, G. Latha, R. Venkatesan, M. Ravichandran, M. A. Atmanand, A. Thirunavukarasu, and G. Raguraman, "Underwater sound to probe sea ice melting in the Arctic during winter," *Scientific Reports*, vol. 10, p. 16047, Dec. 2020.

[10] Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J-N., "ERA5 hourly data on single levels from 1979 to present," 2018.

[11] University of Rhode Island and Inner Space Center, "Audio Gallery." https://dosits.org/galleries/audio-gallery/, June 2017.

[12] NOAA Fisheries, "Hooded Seal." https://www.fisheries.noaa.gov/species/hooded-seal, Wed, 03/23/2022 - 15:25.

[13] U. Send, M. Visbeck, and G. Krahmann, "Aspects of acoustic transponder surveys and acoustic navigation," in *'Challenges of Our Changing Global Environment'. Conference Proceedings. OCEANS '95 MTS/IEEE*, vol. 3, pp. 1631–1642 vol.3, 1995.

[14] Centre national de la recherche scientifique, "CNRS mooring observations at 22°E 81°28'N in the Atlantic Water inflow north of Svalbard - INTAROS Data Catalogue." https://catalog-intaros.nersc.no/dataset/cnrs-mooring-observations-at-22-e-81-28-n-in-the-atlantic-water-inflow-north-of-svalbard, Mar. 2022.

[15] C. Vogt, K. Laihem, and C. Wiebusch, "Speed of sound in bubble-free ice," *The Journal of the Acoustical Society of America*, vol. 124, pp. 3613–3618, Dec. 2008.

Chapter 4

Bayesian optimization with Gaussian process surrogate model for source localization

Source localization with a geoacoustic model requires optimizing the model over a parameter space of range and depth with the objective of matching a predicted sound field to a field measured on an array. We propose a sample-efficient sequential Bayesian optimization strategy that models the objective function as a Gaussian process (GP) surrogate model conditioned on observed data. Using the mean and covariance functions of the GP, a heuristic acquisition function proposes a candidate in parameter space to sample, balancing exploitation (sampling around the best observed objective function value) and exploration (sampling in regions of high variance in the GP). The candidate sample is evaluated, and the GP conditioned on the updated data. Optimization proceeds sequentially until a fixed budget of evaluations is expended. We demonstrate source localization for a shallow-water waveguide using Monte Carlo simulations and experimental data from an acoustic source tow. Compared to grid search and quasi-random sampling strategies, simulations and experimental results indicate the Bayesian optimization strategy converges on optimal solutions rapidly.

4.1 Introduction

Sound propagation in the ocean depends on the physical properties and geometry of the media through which it travels. Using numerical techniques and physical theory, underwater acoustic propagation models predict sound fields in complex environments, enabling estimation of the ocean environment. However, since these models cannot be explicitly inverted, ocean parameters are estimated by sampling parameter space and evaluating samples with a forward model. Forward model predictions are then compared to observed data, with the quality of the prediction given by the correlation—i.e., the match—between the predicted and observed data.

In this study, we present a strategy which casts the inverse problem as a sequential Bayesian optimization problem. Rather than directly optimizing a computationally expensive objective function which could have a complicated structure with many local minima and maxima, a Gaussian process (GP) is fit to the observed data and acts as a surrogate model of the objective function surface. A GP surrogate model is convenient as it provides a tractable approach to modeling the posterior distribution of a function,[1] and has been successfully used to predict sound fields and improve acoustic source direction of arrival estimation, localization, and geoacoustic inversion [2, 3, 4, 5]. More broadly, GP regression is also referred to as kriging and has been used extensively in the geosciences and other fields [6].

The GP model consists of a mean function and covariance function that describe the uncertainty in the objective function: regions which have been sampled exhibit lower uncertainty, while those which remain unexplored exhibit higher uncertainty [1, 7]. Once the GP is fit to the data, a sequential Bayesian framework is applied in which a candidate sample is heuristically proposed by an acquisition function based on the first and second moments of the GP model [8, 9]. The candidate sample is evaluated and the GP model is updated with the new evaluation. This sequential optimization repeats until a maximum budget of objective function evaluations, or trials, is expended.

The design of the acquisition function determines how efficiently the sequential opti-

mization strategy performs [10]. Key to performance of acquisition functions is their ability to balance exploration, sampling in regions of high uncertainty in the GP model, with exploitation, sampling near the best observed value. Conventional acquisition functions derive an analytical function from the mean and covariance functions of the GP model posterior. For example, the upper confidence bound acquisition function proposes a candidate sample where the uncertainty in the GP posterior is greatest [11]. Others, such as the probability of improvement (PI) and expected improvement (EI) acquisition functions, take into account both the uncertainty and mean of the GP posterior, offering an enhanced balance between exploration and exploitation [10]. Recent developments have introduced multi-point versions of conventional acquisition functions, in which the acquisition function suggests multiple candidate samples within each iteration of the sequential optimization [12, 13, 14, 15]. Candidates are suggested through a quasi-Monte Carlo generation scheme or a sequential greedy optimization scheme and tend to outperform conventional acquisition functions in synthetic experiments.

This study demonstrates acoustic source localization in a shallow waveguide using sequential Bayesian optimization with a GP surrogate model. Performance of two acquisition functions is compared to conventional sampling techniques using both acoustic simulations and data from an acoustic source tow experiment. The paper is organized as follows: Section 4.2 summarizes alternative and previous methods; Section 4.3 describes the optimization problem and Bayesian optimization algorithm; Section 4.4 presents simulation and experimental results for source localization in a shallow-water waveguide; and Section 4.5 contains remarks on implementation considerations. Code used for this study is available at https://github.com/NeptuneProjects/BOGP.

4.2 Alternative Strategies for Optimization

Numerous strategies for sampling the parameter space are available. A simple but computationally expensive strategy is matched field processing (MFP), a well known application of grid search typically conducted over source range, depth, and other geoacoustic parameters,

such as ocean bottom composition [16]. A challenge with MFP is extending the parameter search space to more parameters can make the computational cost of the optimization untenable, since the cost of grid search scales exponentially with parameter dimension.

Randomly sampling parameter space can accumulate sufficient evaluations to estimate the global optimum, but is vulnerable to repeatedly sampling from the same region or missing the global optimum altogether [17]. Quasi-random techniques achieve lower discrepancies, ensuring a more even distribution of samples over the parameter space. A popular quasi-random approach uses a Sobol sequence to generate points in high-dimensional space with low discrepancy [18, 19, 20] and has been used for global sensitivity analysis in geoacoustic inversion [21] and in wind turbine noise uncertainty quantification [22].

Grid search, random sampling, and Sobol sampling make no use of the information about the objective function after a sample is evaluated. The gradient of an objective is an example of information that can guide the search. Cases where information about the gradient of the objective is known are often solved with methods such as gradient descent or the Broyden-Fletcher-Goldfarb-Shanno (BFGS) family of optimizers [23]. However, for source localization the ambiguity surface is non-convex and characterized by local optima due to interference patterns. Furthermore, estimating the gradient can be computationally expensive, requiring multiple forward model evaluations. While beamforming methods such as MUSIC [24] and sparse Bayesian learning (SBL)[25, 26, 27, 28] improve resolution and reduce ambiguities in the objective, such methods are still evaluated at grid points and alleviate neither computational cost nor the challenge of non-convexity. Methods such as gradient descent and BFGS are therefore ineffective, as their success depends on the initialization point and might converge on a local optimum.

Recent advances in machine learning have enabled approaches to non-convex optimization such as matrix completion [29, 30] as well as data-driven approaches to acoustic parameter estimation using neural networks and deep learning [31, 32, 33, 34]. Traditional time-difference-of-arrival and bearing-of-arrival localization methods continue to be enhanced through advances

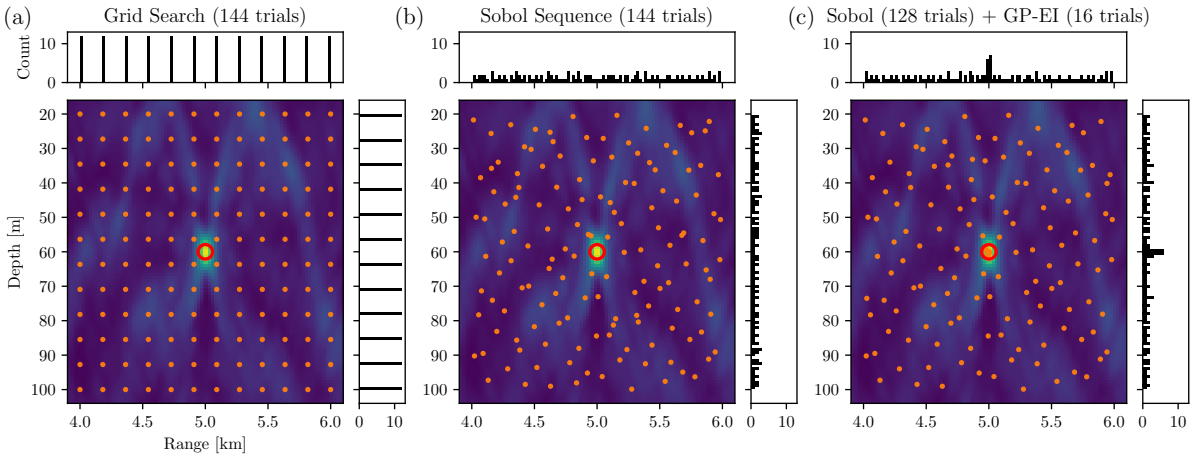

Figure 4.1. (color online) True source location (red circle) and sample locations (orange) for 144 objective function evaluations (trials) using (a) a 12×12 grid search, (b) Sobol sequence sampling, and (c) Bayesian optimization using a Gaussian process with expected improvement acquisition function (GP-EI). Marginal sample histograms are along the axes.

in optimization techniques [35]. In this study, we focus on Bayesian optimization strategies, which use observed values of the objective function to guide the search for the globally optimal solution.

Existing Bayesian approaches to geoacoustic parameter estimation largely rely on treating the ambiguity surface as a posterior distribution over the parameter space and using Monte Carlo sampling to directly estimate the moments of the posterior. A popular approach to geoacoustic parameter estimation uses Markov chain Monte Carlo (MCMC) optimization through genetic algorithms and simulated annealing, with Gibbs sampling providing unbiased moment estimates of the posterior distribution [36, 37, 38, 39, 40]. While simulated annealing and genetic algorithms offer practical and robust methods for global optimization, they take many iterations to converge and require hyperparameter tuning [41, 36, 42, 43]. Numerous advancements in MCMC techniques have since been demonstrated, including trans-dimensional techniques which treat the number of parameters to be estimated as an unknown quantity which must be estimated [44, 45, 46].

To leverage the benefits of local optimization techniques with the global search capa-

bilities of MCMC, hybrid methods combining genetic algorithms with the Gauss-Newton and simplex algorithms have been demonstrated in geoacoustic inversion problems, leading to faster convergence and improved estimation [47, 48]. In a conceptually similar vein, our approach uses quasi-random sampling to initialize sequential Bayesian optimization, hastening convergence and improving optimization performance. Figure 4.1 demonstrates the advantage of sequential Bayesian optimization over grid search and Sobol sampling in an acoustic localization parameter space. Given 144 trials, grid search [Fig. 4.1a] and Sobol sequence sampling [Fig. 4.1b] are unable to resolve the global optimum, whereas Bayesian optimization [Fig. 4.1c] initialized with 128 samples of a Sobol sequence converges on the global optimum within the allotted trial budget.

4.3 Bayesian optimization framework

4.3.1 Objective function definition

Consider an array with M hydrophones that measures an acoustic pressure field $\mathbf{d} \in \mathbb{C}^M$ at a single frequency. In the forward problem, the field measured at the receiver array is described by an acoustic propagation model $G(\mathbf{x})$,

$$\mathbf{d} = G(\mathbf{x}), \tag{4.1}$$

where the parameterization $\mathbf{x} \in \mathscr{X}^D$ describes the acoustic source location and geoacoustic properties of the propagation environment, and $\mathscr{X}^D$ is a domain bounded by finite bounds on each of the D parameters within $\mathbf{x}$. Our task is to solve the inverse problem: given the model G, we seek to find an estimate $\hat{\mathbf{x}}$ of the true parameters $\mathbf{x}_{\text{true}}$ that produce an observed pressure field $\mathbf{d}_{\text{obs}} \in \mathbb{C}^M$.

To find $\hat{\mathbf{x}}$, the parameter space $\mathscr{X}$ is sampled and each sample $\mathbf{x}$ is evaluated using Eq. (4.1), producing replica acoustic pressure fields $\mathbf{d}(\mathbf{x}) \in \mathbb{C}^M$. Relying on the interference patterns that occur due to acoustic propagation in an ocean waveguide, $\hat{\mathbf{x}}$ is obtained by finding

the parameters that yield a predicted replica field which most closely matches the actual field. On every evaluation of $G(\mathbf{x})$, the coherence between the replica field $\mathbf{d}$ and the observed field $\mathbf{d}_{\text{obs}}$ is computed using the Bartlett power:

$$f(\mathbf{x}) = |\mathbf{w}(\mathbf{x})^{\text{H}}\check{\mathbf{d}}|^2, \tag{4.2}$$

where $\mathbf{w}(\mathbf{x})$ is the normalized replica field

$$\mathbf{w}(\mathbf{x}) = \mathbf{d}(\mathbf{x})/\|\mathbf{d}(\mathbf{x})\|; \tag{4.3}$$

and $\check{\mathbf{d}}$ is the normalized observed field

$$\check{\mathbf{d}} = \mathbf{d}_{\text{obs}}/\|\mathbf{d}_{\text{obs}}\| \tag{4.4}$$

to ensure Eq. (4.2) is normalized to the interval $[0, 1]$.

In this localization study, only source range R_{src} and source depth z_{src} are estimated ($\mathbf{x} = [R_{\text{src}}, z_{\text{src}}]^{\mathsf{T}}$); all other geoacoustic properties of the propagation environment are known. We define a multi-frequency objective function by incoherently averaging Eq. (4.2) computed over each frequency in $\Omega = \{\omega_1, \omega_2, \dots\}$ [49, 50]:

$$f(\mathbf{x}) = \frac{1}{|\Omega|}\sum_{i=1}^{|\Omega|} f(\mathbf{x} \mid \omega_i). \tag{4.5}$$

Every evaluation of Eq. (4.5) requires an evaluation of the propagation model at $|\Omega|$ frequencies. Evaluating Eq. (4.5) over $\mathscr{X}$ produces an ambiguity surface whose global maximum occurs where the replica and received pressure fields across all frequencies are most coherent, giving:

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}\in\mathscr{X}}[f(\mathbf{x})]. \tag{4.6}$$

Using the multi-frequency ambiguity surface of Eq. (4.5) to compute Eq. (4.6) improves optimization performance by averaging out frequency-dependent ambiguities, suppressing sidelobes, and smoothing the ambiguity surface, all of which improve the likelihood of converging to the global optimum.

4.3.2 Gaussian process surrogate model

Formally, a GP is a collection of random variables, any finite number of which have a joint Gaussian distribution [7, 1]. Here we follow the derivations of [7]. Consider N samples from the D-dimensional parameter space $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathscr{X}^{D \times N}$. Given a real process $\mathbf{f} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N)]^\mathsf{T}$ as in Eq. (4.5), a GP is described completely by two functions: a mean function,

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{f}] = [\mu(\mathbf{x}_1), \ldots, \mu(\mathbf{x}_N)]^\mathsf{T} \in \mathbb{R}^N, \tag{4.7}$$

where $\mu(\mathbf{x}_n)$ is the mean at $\mathbf{x}_n$; and a covariance function,

$$\begin{aligned} \boldsymbol{\Sigma}_{ij} &= \mathbb{E}[(f(\mathbf{x}_i) - \mu(\mathbf{x}_i))(f(\mathbf{x}_j) - \mu(\mathbf{x}_j))] && (4.8) \\ &= \mathscr{K}(\mathbf{x}_i, \mathbf{x}_j) \in \mathbb{R}^{N \times N}, && (4.9) \end{aligned}$$

where $\mathscr{K}(\mathbf{x}_i, \mathbf{x}_j)$ is a kernel function measuring the similarity between points $\mathbf{x}_i$ and $\mathbf{x}_j$. The GP is summarized as:

$$\mathbf{f} \sim \mathscr{GP}(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{4.10}$$

A GP has observations at sampled parameters comprising the set:

$$\mathscr{D} = \{(\mathbf{x}_n, y_n),\ n = 1 : N\} = \{\mathbf{X}, \mathbf{y}\}, \quad \mathbf{y} \in \mathbb{R}^N \tag{4.11}$$

Though observations of the ambiguity surface are deterministic (in contrast with the data $\mathbf{d}$ used to generate the ambiguity surface), to improve numerical stability in subsequent matrix

inversions, we allow for additive Gaussian noise $\varepsilon_n \sim \mathcal{N}(0, \sigma_y^2)$ in the observations on the order of 10^{-8}:

$$y_n = f(\mathbf{x}_n) + \varepsilon_n. \tag{4.12}$$

Interpolation with a GP is performed by predicting a set of N^* unobserved outputs $\mathbf{f}_*$ at inputs $\mathbf{X}_{*,D\times N^*} = [\mathbf{x}_1^*, \ldots, \mathbf{x}_{N^*}^*]$. The joint distribution of the observed process $\mathbf{y}$ and the predictive distribution $\mathbf{f}_*$ is [7, Eq. (17.33)], [1, Eq. (2.21)]:

$$p(\mathbf{y}, \mathbf{f}_* | \mathbf{X}, \mathbf{X}_*) = \begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_* \end{bmatrix}, \begin{bmatrix} \hat{\mathbf{K}}_{X,X} & \mathbf{K}_{X,*} \\ \mathbf{K}_{X,*}^\mathsf{T} & \mathbf{K}_{*,*} \end{bmatrix} \right) \tag{4.13}$$

where $\boldsymbol{\mu}_X$ and $\boldsymbol{\mu}_*$ are the mean functions at $\mathbf{X}$ and $\mathbf{X}_*$; and

$$\hat{\mathbf{K}}_{X,X} = \mathbf{K}_{X,X} + \sigma_y^2 \mathbf{I} = \mathscr{K}(\mathbf{X}, \mathbf{X})_{N\times N} + \sigma_y^2 \mathbf{I} \tag{4.14}$$

$$\mathbf{K}_{X,*} = \mathscr{K}(\mathbf{X}, \mathbf{X}_*)_{N\times N_*} \tag{4.15}$$

$$\mathbf{K}_{*,*} = \mathscr{K}(\mathbf{X}_*, \mathbf{X}_*)_{N_*\times N_*} \tag{4.16}$$

where $\mathscr{K}$ is a kernel function defined in section 4.3.2. The posterior distribution is obtained by conditioning the GP on the new observations [7, Eq. (17.34)], [1, Eq. (2.22)]:

$$p(\mathbf{f}_* | \mathscr{D}, \mathbf{X}_*) = \mathcal{N}(\mathbf{f}_* | \boldsymbol{\mu}_{*|X}, \boldsymbol{\Sigma}_{*|X}) \tag{4.17}$$

$$\boldsymbol{\mu}_{*|X} = \boldsymbol{\mu}_* + \mathbf{K}_{X,*}^\mathsf{T} \hat{\mathbf{K}}_{X,X}^{-1} (\mathbf{y} - \boldsymbol{\mu}_X) \tag{4.18}$$

$$\boldsymbol{\Sigma}_{*|X} = \mathbf{K}_{*,*} - \mathbf{K}_{X,*}^T \hat{\mathbf{K}}_{X,X}^{-1} \mathbf{K}_{X,*}. \tag{4.19}$$

Kernel function

An important component of the GP surrogate model is the kernel function [Eq. (4.9)], which measures the similarity between two points so that as $\mathbf{x}_i$ and $\mathbf{x}_j$ become more similar, so do their outputs $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$. This relationship is critical to predicting unobserved data points

as in Eqs. (4.13)–(6.18). We use a kernel which is (1) positive definite and (2) stationary with real-valued inputs, i.e.,

$$\mathscr{K}(\mathbf{x}_i, \mathbf{x}_j) = \mathscr{K}(\mathbf{r}), \quad \mathbf{r} = \mathbf{x}_i - \mathbf{x}_j. \tag{4.20}$$

Specifically, we adopt the Matern kernel [1, 7] as it captures variability in length scales and allows for roughness in its output, characteristics which are useful in approximating a non-convex ambiguity surface. The Matern kernel is given in one dimension by:

$$\mathscr{K}(r; \nu, l) = \sigma_y^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} r}{l} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu} r}{l} \right) \tag{4.21}$$

where K_ν is the modified Bessel function. $\sigma_y \in \mathbb{R}$, $l \in \mathbb{R}$, and $\nu \in \{1/2, 3/2, 5/2, \dots\}$ are hyperparameters, with σ_y^2 estimating the noise variance of the GP, l controlling the length scale, and ν controlling the roughness of the kernel output. Higher values of ν result in smoother outputs, with $\nu \to \infty$ giving the squared exponential kernel; we adopt the typical choice of $\nu = 5/2$ [7, Eq. (17.13)], [1, Eq. (4.17)]. Using automatic relevance determination (ARD), characteristic length scales are estimated for each dimension by modifying Eq. (4.21) to [51, Sec. 1.2.3], [1, Sec. 5.1], [7, Eq. (17.8)]:

$$\mathscr{K}\left(\mathbf{r}; \frac{5}{2}, \mathbf{l}\right) = \sigma_y^2 \prod_{d=1}^{D} \left(1 + \frac{\sqrt{5} r_d}{l_d} + \frac{5 r_d^2}{3 l_d^2} \right) \exp\left(-\frac{\sqrt{5} r_d}{l_d} \right), \tag{4.22}$$

where r_d is the distance between points $\mathbf{x}_1$ and $\mathbf{x}_2$ along dimension d, and characteristic length scales for each dimension are given by $\mathbf{l} = [l_1, l_2, \dots, l_D]$.

Kernel hyperparameter optimization

The Matern kernel function contains hyperparameters $\boldsymbol{\theta} = [\sigma_y^2, \mathbf{l}]$ which must be optimized for the GP surrogate model to appropriately reflect the data [1, 7]. Kernel function fitting occurs after new samples are drawn but before the predictive distribution is computed in Eq. (4.13) and is efficiently solved with an empirical-Bayes approach [7]. To optimize $\boldsymbol{\theta}$, the

marginal likelihood is maximized [7, Eq. (17.51)]:

$$p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f},\mathbf{X})p(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})d\mathbf{f}. \tag{4.23}$$

where, treating $\boldsymbol{\theta}$ as implicit in $\hat{\mathbf{K}}_{X,X}$:

$$p(\mathbf{f}|\mathbf{X}) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}_X, \hat{\mathbf{K}}_{X,X}) \tag{4.24}$$

$$p(\mathbf{y}|\mathbf{f}) = \prod_{n=1}^{N} \mathcal{N}(y_n|f_n, \sigma_y^2). \tag{4.25}$$

The log marginal likelihood and its derivative are then [52, Eq. (18.74)], [7, Eq. (17.52)]:

$$\begin{aligned} L &= \log p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}) = \log \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_X, \hat{\mathbf{K}}_{X,X}) \\ &= -\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_X)^\mathsf{T} \hat{\mathbf{K}}_{X,X}^{-1}(\mathbf{y}-\boldsymbol{\mu}_X) - \frac{1}{2}\log|\hat{\mathbf{K}}_{X,X}| - \frac{N}{2}\log(2\pi) \end{aligned} \tag{4.26}$$

$$\frac{\partial L}{\partial \theta_j} = \frac{1}{2}\mathrm{tr}\left[(\boldsymbol{\alpha}\boldsymbol{\alpha}^\mathsf{T} - \hat{\mathbf{K}}_{X,X})\frac{\partial \hat{\mathbf{K}}_{X,X}}{\partial \theta_j}\right], \tag{4.27}$$

where $\boldsymbol{\alpha} = \hat{\mathbf{K}}_{X,X}^{-1}(\mathbf{y}-\boldsymbol{\mu}_X)$. In contrast to the ambiguity surface, since Eq. (4.26) is smoothly varying with few optima, hyperparameter optimization is performed with the bounded Limited-memory BFGS (L-BFGS-B) algorithm, a quasi-Newtonian method that finds a minimizing solution given a starting point $\mathbf{x} \in \mathscr{X}$ and a smooth objective function [53, 54]. L-BFGS-B terminates either by comparing Eq. (4.26) between iterations k and $k+1$ according to the condition:

$$\frac{L^k - L^{k+1}}{\max\left(|L^k|, |L^{k+1}|, 1\right)} \leq 10^7 \varepsilon \tag{4.28}$$

where $\varepsilon \sim \mathscr{O}(10^{-16})$ is the machine precision; or when Eq. (4.27) projected onto the feasible parameter space $\mathscr{X}$ (denoted by the Proj operator) meets the condition:

$$\max_{j\in\{1,\ldots,D\}} \left|\mathrm{Proj}\left(\frac{\partial L}{\partial \theta_j}\right)\right| \leq 10^{-5}. \tag{4.29}$$

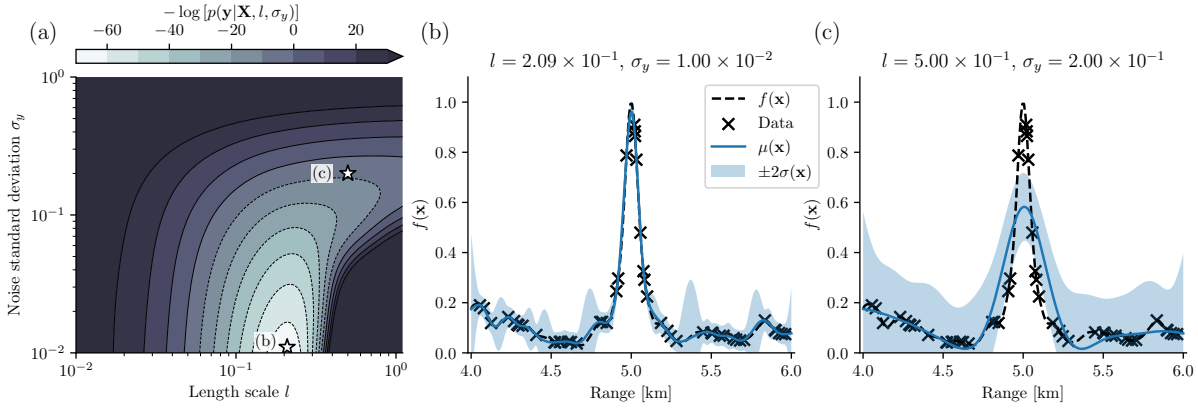

Figure 4.2. Hyperparameter optimization for Gaussian process (GP) regression on a one-dimensional broadband ambiguity surface computed over source range. (a) Negative log-likelihood of a Matern kernel function vs. the noise standard deviation and length scale hyperparameters. Labeled stars indicate the resulting GP regression for (b) the optimal fit and (c) a suboptimal fit.

To improve convergence in hyperparameter optimization, prior distributions are placed over the kernel hyperparameters from which starting points for L-BFGS-B are selected. For computational stability, parameters are normalized to $[0,1]$ and observations standardized to zero mean and unit variance before fitting the GP. For length scales $\mathbf{l}$, a Gamma distribution [7, Sec. 2.7.5] is adopted with shape $a = 3$ and rate $b = 6$, yielding a distribution with a mean of 0.5. We find this is a reasonable choice for the transformed parameter space as the prior encourages samples consistent with expected correlation length scales given the source frequencies and geometry of the shallow waveguide [55]. For the noise variance σ_y^2, a Gamma distribution with shape $a = 2$ and rate $b = 0.15$ is adopted.

Examples of Matern kernel hyperparameter optimization are shown in Fig. 4.2 for a one-dimensional GP regression on a broadband ambiguity surface computed over source range. Figure 4.2a shows the negative log likelihood [Eq. (4.26)] over the $\boldsymbol{\theta} = [\sigma_y^2, l]$ hyperparameters. The model fits indicated by the white stars correspond to the optimal GP regression in Fig. 4.2b and a suboptimal GP regression in Fig. 4.2c whose noise variance and length scale are too large.

4.3.3 Acquisition functions

Acquisition functions provide a heuristic which guides the sequential sampling strategy. Since the dataset $\mathscr{D}$ grows with every iteration, and to differentiate from the sample index n, we introduce a trial index t to denote sequential operations. At every trial t, a new sample is drawn, evaluated, and appended to the previously evaluated data by:

$$\mathscr{D}|_t = \mathscr{D}|_{t-1} \cup \{(\mathbf{x}_t, f(\mathbf{x}_t))\}. \tag{4.30}$$

To select the next point in parameter space $\mathbf{x}_t$ which will be evaluated, an acquisition function α takes the GP predictive posterior distribution as its input and returns a proposed candidate for the next trial by:

$$\mathbf{x}_t = \arg\max_{\mathbf{x}} \alpha(f(\mathbf{x})). \tag{4.31}$$

Numerous algorithmic implementations are available to compute α; here we evaluate two heuristics defined by the expected improvement over previous observations. Given a set of observations $\mathscr{D}$, let f' be the largest observed value of $\mathbf{f}$ and the improvement over f' at any point $\mathbf{x}$ be defined as[10]:

$$I(\mathbf{x}) = \max(f(\mathbf{x}) - f', 0), \tag{4.32}$$

noting that $I \geq 0$. I is a random variable since the uncertainty of the objective function is encoded in $\mathbf{f} \sim \mathscr{N}(\boldsymbol{\mu}, \blacksquare)$.

The covariance function $\blacksquare$ of the posterior distribution $\mathbf{f}$ is assumed diagonal and the variance $\sigma^2(\mathbf{x})$ is:

$$\boldsymbol{\Sigma} \approx \operatorname{diag}\left[\sigma^2(\mathbf{x}_1), \ldots, \sigma^2(\mathbf{x}_N)\right]. \tag{4.33}$$

Using $\sigma(\mathbf{x})$, we use the re-parameterization trick to rewrite the posterior distribution as:

$$f(\mathbf{x}) = \mu(\mathbf{x}) + \sigma(\mathbf{x})z. \tag{4.34}$$

Table 4.1. Optimization of analytic (Part A, EI and PI) and quasi-Monte Carlo (Part B, qEI) acquisition functions.

Input: Parameter domain $\mathscr{X}$, acquisition function α, number of raw samples N_{raw}, number of restarts N_{restart}
Output: Next sample point $\mathbf{x}_t$ or points $\mathbf{X}_t$
Initialization:
$\mathbf{x}_t \leftarrow \mathbf{0}$, $\check{\alpha} \leftarrow 0$
$\tilde{\mathbf{X}} \leftarrow$ Draw N_{raw} i.i.d. samples from $\mathscr{X}$
$\tilde{\boldsymbol{\alpha}} \leftarrow \alpha(\tilde{\mathbf{X}})$ [Eq. (4.42)]
$\mathbf{z} \leftarrow \frac{\tilde{\boldsymbol{\alpha}} - \text{mean}(\tilde{\boldsymbol{\alpha}})}{\text{std}(\tilde{\boldsymbol{\alpha}})}$ [Eq. (4.43)]
Part A: Expected Improvement
for $i = 1$ **to** N_{restart} **do**
  $\mathbf{x} \leftarrow$ Draw sample from $p(e^{\mathbf{z}})$
  $\check{\mathbf{x}} \leftarrow \text{L-BFGS-B}\left[-\alpha(\mathbf{x})\right]$
  if $\alpha(\check{\mathbf{x}}) > \check{\alpha}$ **then**
    $\check{\alpha} \leftarrow \alpha(\check{\mathbf{x}})$
    $\mathbf{x}_t \leftarrow \check{\mathbf{x}}$
Part B: q-Expected Improvement
for $i = 1$ **to** N_{restart} **do**
  $\mathbf{x} \leftarrow$ Draw sample from $p(e^{\mathbf{z}})$
  for $j = 1$ **to** q **do**
    $\check{\mathbf{x}} \leftarrow \text{L-BFGS-B}\left[-\alpha(\mathbf{x})\right]$
    $\text{col}_j\left[\check{\mathbf{X}}\right] \leftarrow \check{\mathbf{x}}$
    $\mathbf{x} \leftarrow \check{\mathbf{x}}$
  if $\alpha\left(\text{col}_{N_{\text{restart}}}\left[\check{\mathbf{X}}\right]\right) > \check{\alpha}$ **then**
    $\check{\alpha} \leftarrow \alpha\left(\text{col}_{N_{\text{restart}}}\left[\check{\mathbf{X}}\right]\right)$
    $\mathbf{X}_t \leftarrow \check{\mathbf{X}}$

The improvement as defined in Eq. (4.32) is then rewritten:

$$I(\mathbf{x}) = \max(\mu(\mathbf{x}) + \sigma(\mathbf{x})z - f', 0). \tag{4.35}$$

Expected improvement (EI)

Expected improvement (EI) gives the expected magnitude in improvement over the best previously observed evaluation of the objective function [10]. Defining z_0 as:

$$z_0(\mathbf{x}) = \frac{f' - \mu(\mathbf{x})}{\sigma(\mathbf{x})}, \tag{4.36}$$

EI is computed by evaluating the upper side of the cumulative distribution function $\Phi(z)$ and the normal probability distribution function $\varphi(z)$:

$$\begin{aligned} \mathrm{EI}(\mathbf{x}) &= \mathbb{E}[I(\mathbf{x})] = \mathbb{E}[\max(\mu(\mathbf{x}) + \sigma(\mathbf{x})z - f', 0)] \\ &= \left(\mu(\mathbf{x}) - f'\right)\left(1 - \Phi(z_0(\mathbf{x}))\right) + \sigma(\mathbf{x})\varphi(z_0(\mathbf{x})) \end{aligned} \tag{4.37}$$

EI is increased either by reduction of the mean $\mu(\mathbf{x})$ (exploitation) or the variance $\sigma(\mathbf{x})$ (exploration). From Eq. (4.31), the point $\mathbf{x}_t$ that maximizes Eq. (4.37) is selected for evaluation in the subsequent trial.

The EI acquisition function assumes a noise free measurement of f', but from Eq. (4.12), a small amount of noise in the objective function evaluation is expected. To avoid slow or incorrect convergence, we use an implementation of EI which accounts for a noisy objective function [56].

q-Expected Improvement (qEI)

Recent work [12, 13, 14, 15] extends improvement-based acquisition functions by implementing parallel evaluations of the acquisition function. One such implementation, quasi-Monte Carlo EI (q-Expected Improvement, or qEI), evaluates a batch of q random samples $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_q] \in \mathscr{X}$ with Eq. (4.37). The q resulting values of EI are averaged, giving qEI for the q-batch:

$$\mathrm{qEI}(\mathbf{X}) = \mathbb{E}\left[\max_{i=1,\ldots,q}\left(\mu(\mathbf{x}_i) + \sigma(\mathbf{x}_i)z - f', 0\right)\right] \tag{4.38}$$

$$\approx \frac{1}{q}\sum_{i=1}^{q}\mathrm{EI}(\mathbf{x}_i) \tag{4.39}$$

Unlike the analytical acquisition function EI which yields only one candidate, qEI yields q candidates to be subsequently evaluated in parallel, modifying Eqs. (4.31) and (4.30) to:

$$\mathbf{X}_t = \underset{\mathbf{X}=[\mathbf{x}_1,\ldots,\mathbf{x}_q]}{\arg\max}\ \alpha(f(\mathbf{X})) \tag{4.40}$$

$$\mathscr{D}|_t = \mathscr{D}|_{t-1} \cup \{(\mathbf{X}_t, f(\mathbf{X}_t))\}. \tag{4.41}$$

qEI therefore samples the acquisition function and objective function spaces more rapidly than the analytical acquisition functions. However, as q increases, the optimization shifts from a Bayesian framework to a Monte Carlo framework. In this study, $q = 4$ was found to appropriately balance the benefits of sequential Bayesian optimization with the robustness of quasi-Monte Carlo sampling.

Acquisition function optimization

Exhaustively evaluating the acquisition function to solve Eq. (4.31) or Eq. (4.40) is computationally expensive for higher dimensional parameter spaces. In practice, an auxiliary optimization is performed to suggest samples for the next trial. Here again the L-BFGS-B algorithm is used [53, 54], and as Eqs. (4.31) and (4.40) are maximization problems, the auxiliary optimization is transformed to a minimization problem by supplying the negated acquisition function as the objective function.

Since acquisition functions are often non-convex and can contain large regions with zero gradient, the auxiliary optimization is sensitive to the starting point. To improve performance, we rely on heuristics whereby the auxiliary optimization is performed using N_{restart} restarts. Starting points are chosen by first evaluating the acquisition function α at N_{raw} random points $\tilde{\mathbf{X}} \in \mathscr{X}$:

$$\tilde{\boldsymbol{\alpha}} = \alpha(\tilde{\mathbf{X}}). \tag{4.42}$$

From these random evaluations of the acquisition function, a sample Z-distribution is computed:

$$\mathbf{z} = \frac{\tilde{\boldsymbol{\alpha}} - \text{mean}(\tilde{\boldsymbol{\alpha}})}{\text{std}(\tilde{\boldsymbol{\alpha}})} \tag{4.43}$$

which is used to construct an exponentiated distribution $p(e^{\mathbf{z}})$ from which N_{restart} starting points are drawn without replacement. The acquisition function's maximizer $\check{\mathbf{x}}$ is returned by L-BFGS-B for each of the N_{restart} starting points. The point $\check{\mathbf{x}}$ corresponding to the highest value of $\alpha(\check{\mathbf{x}})$ from the N_{restart} restarts provides the next candidate point $\mathbf{x}_t$. Table 4.1 Part A summarizes optimization of the EI acquisition function.

For each of the N_{restart} auxiliary optimizations for the quasi-Monte Carlo acquisition function qEI, q samples $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_q] \in \mathscr{X}$ are sequentially drawn from $p(e^{\mathbf{z}})$. Each of the q samples serves as a starting point for maximizing the acquisition function by L-BFGS-B, and the maximizing points $\check{\mathbf{X}}$ are evaluated by the acquisition function. The maximizing points $\check{\mathbf{X}}$ corresponding to the highest value of $\alpha(\check{\mathbf{X}})$ from the N_{restart} restarts provides the next candidate points $\mathbf{X}_t$. Table 4.1 Part B summarizes optimization of the qEI acquisition function.

4.3.4 Implementation

Implementation of Bayesian optimization with a GP surrogate model is shown in Table 4.2, and parameter values used are in Table 4.3. First, N_{init} samples are drawn from the parameter space using a Sobol sequence[20] and evaluated by the objective function. The initializing trials establish a reasonable prior distribution for the GP surrogate model. With the GP model fit, the Bayesian framework proceeds by optimizing the acquisition function α and generating candidate samples which are evaluated. Data from the new trials are appended to the existing data $\mathscr{D}$, and the GP model is fit with the updated data. This process repeats until the total number of trials N has been reached.

Bayesian optimization using the EI acquisition function is illustrated in Fig. 4.3 for one-dimensional range estimation for a simulated broadband source at 5.0 km; environment

Table 4.2. Bayesian optimization with GP surrogate model.

Input: Parameter domain $\mathscr{X}$, objective function f, kernel function $\mathscr{K}$, acquisition function α, warmup evaluations N_{init}, total evaluations N
Output: Best estimate of parameters $\hat{\mathbf{x}}$
Initialization:
$f' \leftarrow 0$, $\boldsymbol{\theta} \leftarrow \mathbb{R}$
for $t = 1$ **to** N_{init} **do**
  $\mathbf{x}_t \leftarrow \text{SOBOL}[\mathscr{X}]$ [20]
  $f_t \leftarrow f(\mathbf{x}_t)$ [Eq. (4.5)]
  if $f_t > f'$ **then**
    $\hat{\mathbf{x}} \leftarrow \mathbf{x}_t$, $f' \leftarrow f_t$
Optimization:
for $t = N_{\text{init}} + 1$ **to** N **do**
  $\boldsymbol{\mu} \leftarrow \mathbb{E}[f(\mathbf{X})]$ [Eq. (4.7)]
  $\boldsymbol{\theta} \leftarrow \text{L-BFGS-B}\,[\mathscr{K}(\mathbf{X}, \mathbf{X}; \boldsymbol{\theta})]$ [Section 4.3.2]
  $\boldsymbol{\Sigma} \leftarrow \mathscr{K}(\mathbf{X}, \mathbf{X}; \boldsymbol{\theta})$ [Eq. (4.9)]
  $\mathscr{GP} \leftarrow \mathscr{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ [Eq. (4.10)]
  $\mathbf{x}_t \leftarrow \arg\max_{\mathbf{x}} \alpha(\mathbf{x})$ [Algorithm 4.1]
  $f_t \leftarrow f(\mathbf{x}_t)$
  if $f_t > f'$ **then**
    $\hat{\mathbf{x}} \leftarrow \mathbf{x}_t$, $f' \leftarrow f_t$

Table 4.3. Bayesian optimization strategy parameters.

Parameter	Description	Value
N	Total trials	144
N_{init}	Warm-up trials	128
N_{restart}	Acquisition function re-starts	40
N_{raw}	Raw samples for acquisition function optimization [Table 4.1]	1024
q	Batch size for q-Expected Improvement	4

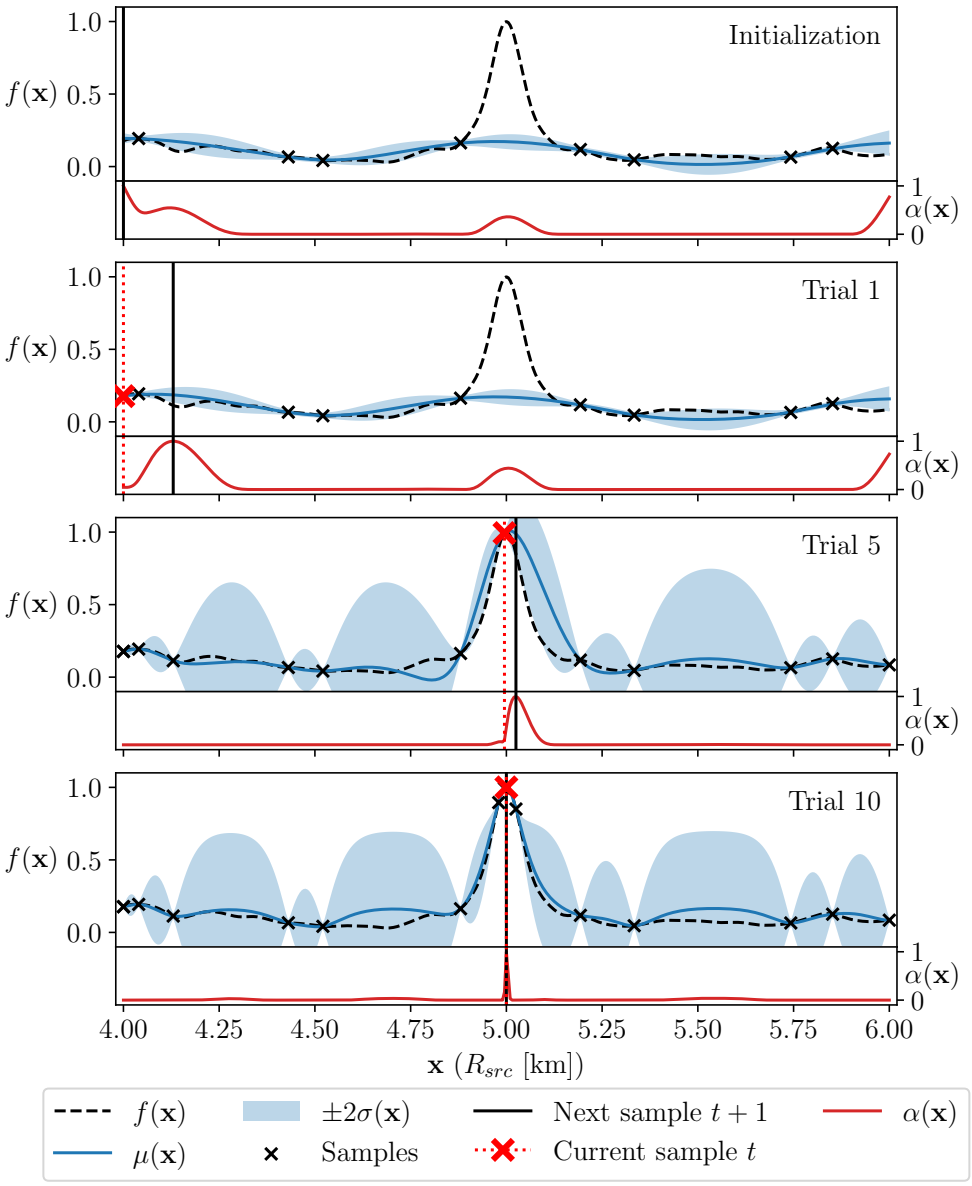

Figure 4.3. (color online) Range estimation for a simulated broadband source at 60 m depth and 5 km range using Bayesian optimization with GP surrogate model. Optimization is initialized with eight quasi-random samples. Top panels show the true objective function $f(\mathbf{x})$ (black dashed), and the mean function (blue) and standard error (blue shaded) of the GP. Bottom panels show the normalized expected improvement acquisition function $\alpha(\mathbf{x})$ [Eq. (4.37)]. The maximum of the acquisition function (vertical solid line) guides the location of the subsequent trial.

and array details are given in Section 4.4.1. The GP is initialized with 8 trials sampled by a Sobol sequence. GP fit to the data is poor in early iterations but improves as observations are added. The acquisition function adaptively balances exploitation and exploration, with maximum values alternating between areas of high uncertainty (exploration) and areas with large objective function values (exploitation). The optimal solution is generally located by the 5^{th} trial; in later trials, the acquisition function shape approaches a delta function, indicating a high confidence in the solution.

4.4 Results

Source localization is demonstrated using simulated and experimental data. A fixed budget of 144 trials is set for all optimization strategies. For the Bayesian strategies, the GP prior distribution is initialized with 128 trials sampled by a Sobol sequence. Two Bayesian strategies are evaluated: Sobol+GP/EI, which uses the EI acquisition function, and Sobol+GP/qEI, which uses the qEI acquisition function with $q = 4$. For comparison, 144 trials of Sobol sequence sampling are evaluated, as well as 144 trials of grid search on an evenly spaced 12×12 grid.

4.4.1 Simulations

Measured data $\mathbf{d}$ are simulated from the SWellEx-96 experiment [57, 58]. The environment, represented in Fig. 4.4, comprises a shallow, downward-refracting waveguide. A vertical line array (VLA) of 64 hydrophones evenly spaced between depths of 94.125 and 212.5 m (1.875 m spacing) recorded signals from an acoustic source towed by RV *Sproul* at a depth of approximately 60 m. The measured acoustic field $\mathbf{d}$ and replica fields $\mathbf{d}^*$ are computed using the KRAKEN normal mode propagation model [59]. The source is simulated transmitting at $\Omega = \{148, 166, 201, 235, 283, 338, 388\}$ Hz from depth $z_{\text{src}} = 60$ m at ranges $R_{\text{src}} = \{1.0, 3.0, 5.0, 7.0\}$ km. The range search space is a ± 1 km window centered around the simulated source ranges R_{src}. For $R_{\text{src}} = 1.0$ km, the range search space is 1.01 ± 1 km to avoid near field effects. The depth search space is 60 ± 40 m. Because the Bayesian and Sobol

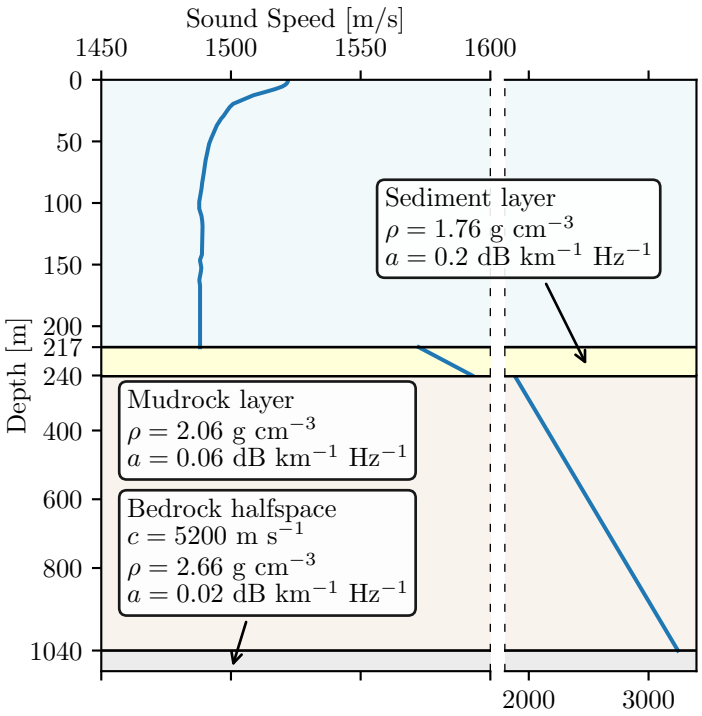

Figure 4.4. Sound speed profile and geoacoustic properties used for simulating acoustic propagation.

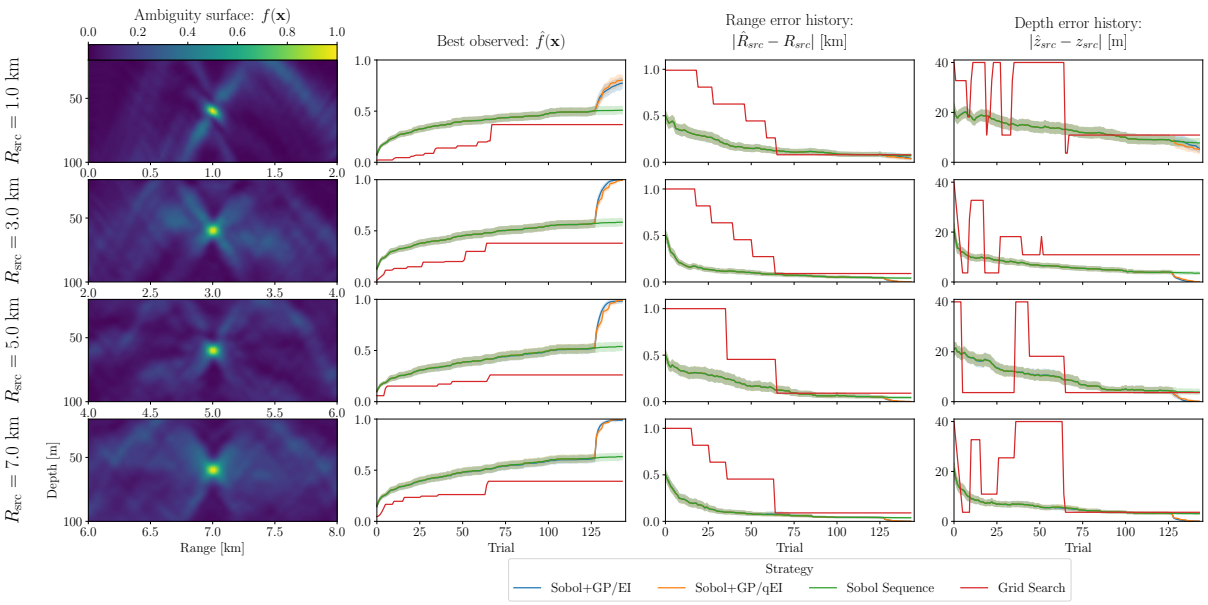

Figure 4.5. (color online) Two-dimensional matched field processing (MFP) multi-frequency ambiguity surfaces (left column) for a simulated source at 60 m depth and 1, 3, 5, and 7 km range. Best observed optimization performance (middle-left column), source range error (middle-right column), and source depth error (right column) from 100 Monte Carlo simulations are shown for each trial. Solid colored lines indicate mean values and shaded regions indicate standard deviation.

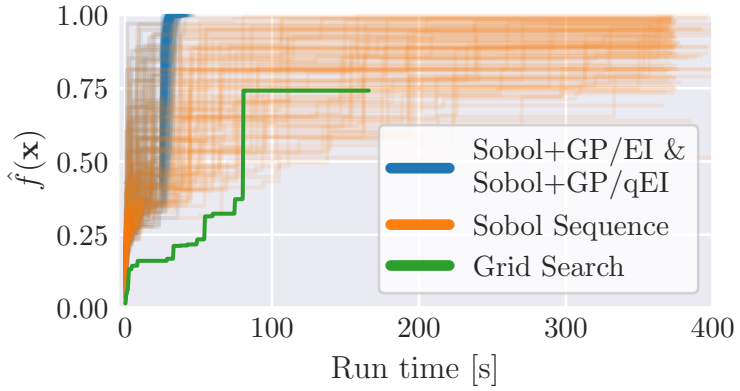

Figure 4.6. (color online) Highest observed objective vs. run time using a simulated source at $R_{\mathrm{src}} = 3.0$ km and $z_{\mathrm{src}} = 60$ m. Sobol+GP/EI and Sobol+GP/qEI (blue) consist of 128 Sobol sequence trials followed by 16 GP/EI or GP/qEI steps (144 total trials); Sobol sequence (orange) of 1,024 trials; and grid search (green) of a 32×32 grid (1,024 trials).

sequence strategies are quasi-random, 100 Monte Carlo simulations are performed to assess performance.

Figure 4.5 shows localization results for each optimization strategy, with each row corresponding to a simulated source range. The two-dimensional broadband ambiguity surface of Eq. (4.5) is shown in the left column and constitutes the objective function $f(\mathbf{x})$ modeled by the GP surrogate model. In all cases, Sobol+GP/EI and Sobol+GP/qEI outperform Sobol sampling and grid search according to the best observed objective function value $\hat{f}$, range error, and depth error. This is due to the grid search evaluating local optima. Grid search performance is noteworthy in that, though $\hat{f}$ increases as the grid is evaluated, the apparent improvement might not give a decrease in range and depth errors.

Though fitting the GP surrogate model and optimizing the acquisition function are somewhat computationally expensive, on average the Bayesian strategies converge on the global optimum in far fewer trials than the Sobol sequence and grid search strategies alone. Figure 4.6 shows traces of $\hat{f}$ as a function of run time for the 100 Monte Carlo runs of each strategy. Run times were measured one run at a time on a 16-core laptop computer. The Bayesian strategies ran for 144 trials (128 Sobol, 16 Bayesian), while the Sobol sequence and grid search strategies ran for 1,024 trials. More trials and time are required for Sobol sampling and grid search to converge on the global optimum than for the Bayesian strategies.

4.4.2 Experimental Data

The following analysis uses SWellEx-96 experimental data recorded during event S5, conducted between 23:15-00:30 GMT on 10-11 May 1996 12 km west of Point Loma, California [57, 58]. RV *Sproul* towed an acoustic source along an isobath of 200 m, though actual depth varied, with the source tow commencing in 220 m of water and the second half of the tow occurring in 180 m of water. RV *Sproul* proceeded from south to north at 5 knots (2.5 m/s), with a closest point of approach (CPA) to the VLA of 900 m. The source was towed at 60 m depth and transmitted five sets of tonals with varying source levels. Data were recorded by the VLA

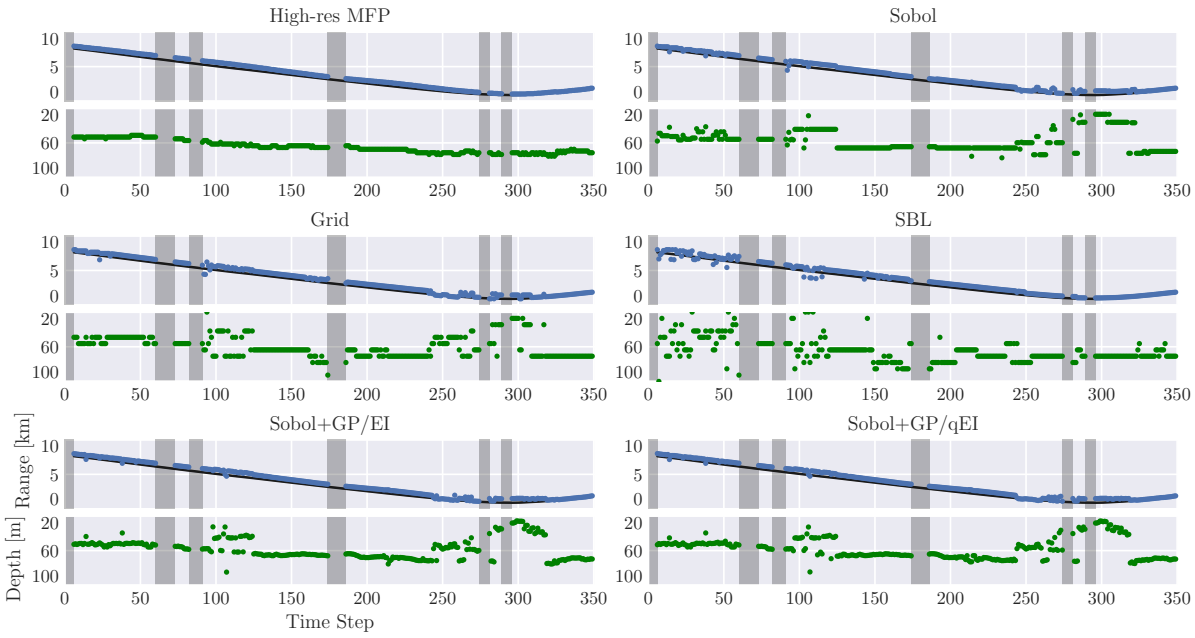

Figure 4.7. (color online) Range (blue) and depth (green) estimated localization for high-resolution matched field processing (MFP), low-resolution MFP (grid search), sparse Bayesian learning grid search (SBL), Sobol sampling, and Bayesian optimization using expected improvement (Sobol+GP/EI) and quasi-Monte Carlo expected improvement (Sobol+GP/qEI) acquisition functions. The black line indicates the GPS range of RV *Sproul* to the array. Gray shaded areas indicate when the source stopped transmitting.

with a 1500-Hz sampling rate. The 64-element array is described in Section 4.4.1.

Data between 23:21-00:24 GMT were processed in 350 non-overlapping time steps. At the starting point, half-way point, and CPA of the source tow, the deep source ceased broadcasting CW tonals and transmitted frequency-modulated (FM) chirps; these time segments are omitted.

Replica vectors $\mathbf{d}^*$ are calculated at each frequency in $\Omega = \{148, 166, 201, 235, 283, 338, 388\}$ Hz using the environmental model in Fig. 4.4. These frequencies correspond to the upper seven tonals from the the loudest set transmitted by the source. To approximate array tilt and improve localization, a 1° tilt away from the source is applied to replica vector calculations at all time steps [58]. Measurement vectors $\mathbf{d}$ are obtained from the discrete Fourier transform of the experimental data at each time step and at the frequencies in Ω.

The range search space is a ± 1 km window centered around the GPS position of

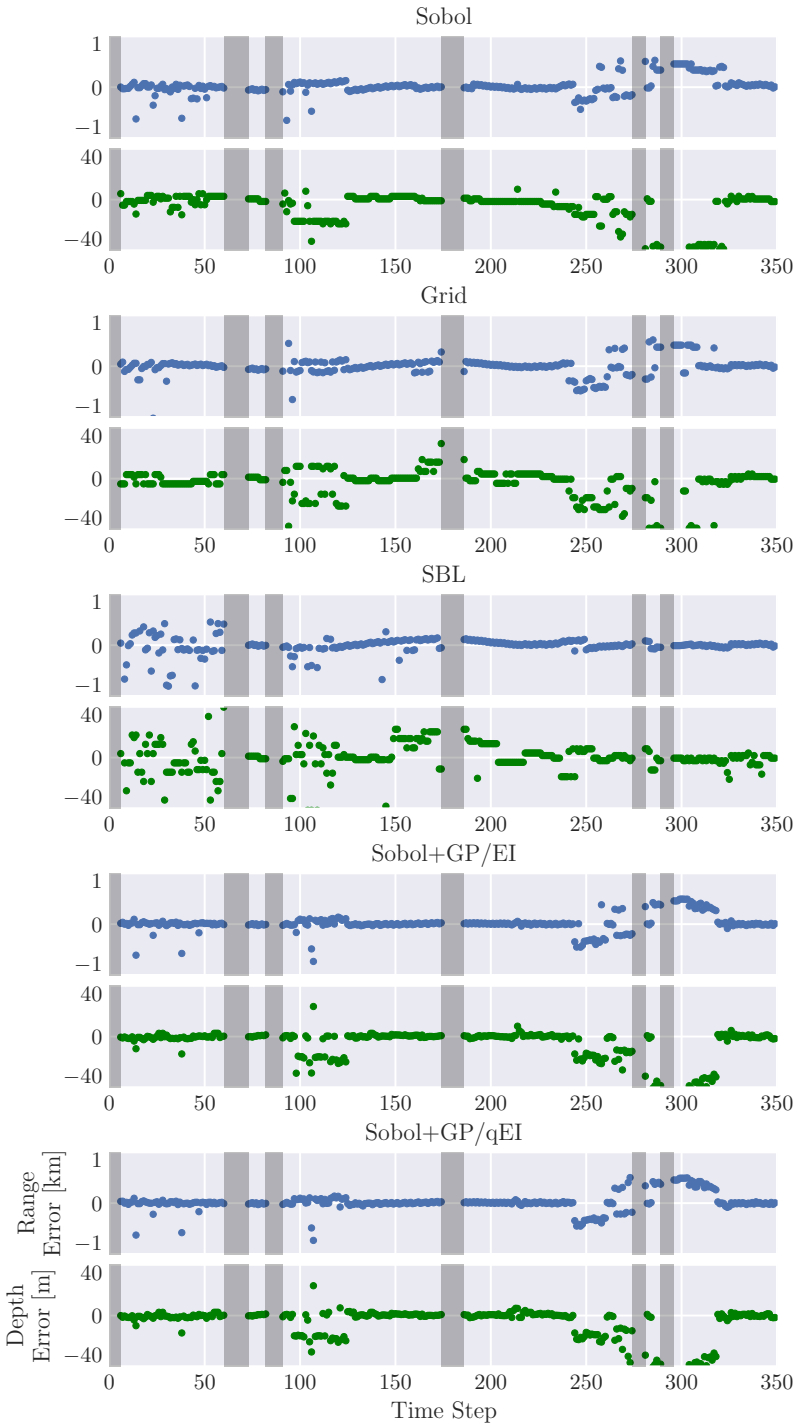

Figure 4.8. (color online) Range (blue) and depth (green) estimation errors relative to high-resolution matched field processing. Gray shaded areas indicate when the source stopped transmitting.

Table 4.4. Mean absolute error (MAE) and median absolute error (Med AE) with respect to high-resolution matched field processing.

	Range [km]		Depth [m]	
Strategy	MAE	Med AE	MAE	Med AE
Grid	0.111	0.046	7.795	3.576
SBL	0.130	0.048	7.500	3.697
Sobol	0.107	0.041	8.308	2.543
Sobol+GP/EI	**0.090**	**0.017**	**7.125**	**0.974**
Sobol+GP/qEI	0.093	0.017	7.477	1.286

RV *Sproul* (R_{GPS}) at each time step. When $R_{\text{GPS}} < 1.01$ km, the range search space is 1.01 ± 1 km to avoid near field effects; when $R_{\text{GPS}} > 7.0$ km, the search space is 7.0 ± 1 km. The depth search space is 60 ± 40 m.

High-resolution MFP (200 range bins, 100 depth bins) establishes a baseline against which to compare optimization strategies and compute localization error. Results from high-resolution MFP are plotted in the upper left panel of Fig. 4.7 for each time step. As RV *Sproul* approaches the array, the ship's GPS range is closer than the high-resolution MFP estimate due to the scope of the cable towing the source; at CPA, the disparity expectedly reverses. Over the course of the source tow, the high-resolution MFP depth estimate indicates a gradual depth change from 50 to 70 m. The apparent depth change results from a mirage effect arising from mismatch between true and modeled (217 m) bathymetry used to compute replica vectors $\mathbf{d}^*$ [60].

Range and depth estimation are performed using the Sobol+GP/EI and GP/qEI Bayesian optimization strategies, and the grid search and Sobol sampling as described in Section 4.4.1. For additional comparison, results from SBL [25, 26, 27, 28, 61] computed at the same points as the grid search are presented. No prior information is passed from one time step to the next, and the optimizations are reinitialized at each time step. Figure 4.7 shows results of range and depth estimation at each time step for all strategies, and Fig. 4.8 shows the range and depth estimation errors relative to high-resolution MFP. All methods localize the source reasonably well, but the

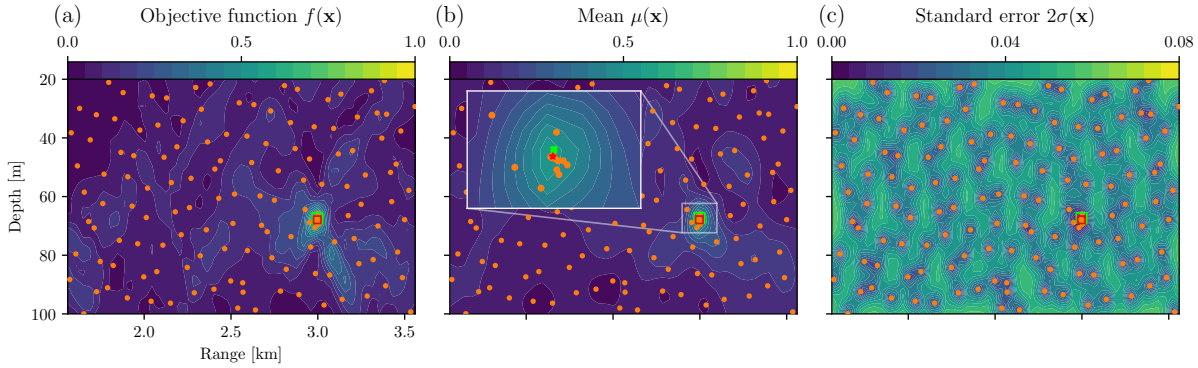

Figure 4.9. (color online) (a) Objective function (ambiguity surface), (b) mean function, and (c) standard error surface for the GP posterior at time step 200 ($R_{\text{GPS}} = 2.56$ km). Optimization was performed using the EI acquisition function. Samples (orange circles) and the actual (green) and estimated (red) source positions are indicated. The inset in (b) shows the dense sampling pattern and best estimate from Bayesian optimization converging on the global optimum.

Bayesian methods are able to track the source more closely than grid search and Sobol sampling. At longer ranges, Bayesian methods track the source more closely than SBL, whereas at close range, SBL out-performs the Bayesian methods. The mean absolute and median absolute errors over the entire source tow are listed for each strategy in Table 4.4. Except for SBL, performance suffers when the source is near CPA, likely due to the complicated structure of the ambiguity surface at this range: from Fig. 4.5, the ambiguity surface at $R_{\text{src}} = 1.0$ km has local optima in close proximity to the global optimum.

Figure 4.9 compares the GP mean and standard error surfaces to the objective function surface upon completion of the Sobol+GP/EI strategy for time step 200. In the region surrounding the global optimum where sampling is most dense, the mean function $\mu(\mathbf{x})$ resembles the objective function $f(\mathbf{x})$ and has low variance.

4.5 Discussion

We find the success of the Sobol+GP/EI and Sobol+GP/qEI strategies rests on establishing a reasonable prior over the objective function prior to commencing Bayesian optimization. Two primary factors contribute to this prior: the domain of the parameter space and the number

of warm-up trials.

The complicated structure of the ambiguity surface informed our choice of a parameter space constrained to 2 km range and 80 m depth windows. When a broader search space is used (e.g., 1 to 10 km range and 0 to 200 m depth), there are more local optima in the parameter space and Bayesian optimization converges on the optimal solution less reliably. This can be mitigated by using more warm-up trials to obtain a better prior.

The appropriate balance between warm-up trials and Bayesian optimization trials is a design consideration which must be evaluated according to the data, objective function, parameter space, and computational budget available. We arrived at a ratio of 128 warm-up trials to 16 Bayesian optimization trials through experimentation. Initial attempts to invert the ratio resulted in a poor prior, and Bayesian optimization was unable to reliably converge on a solution due to the complicated structure of the objective function. Since Bayesian optimization trials take approximately 1 s to fit the GP and optimize the acquisition function, using 16 warm-up trials and 128 Bayesian trials not only resulted in poor convergence but also expended more computation time. The ratio of warm-up trials to Bayesian optimization trials is therefore best evaluated against the expected number of optima in the objective function, which is itself dependent on the domain of the parameter space.

Experimentation with adjusting the size of the parameter space and the ratio of warm-up trials to Bayesian optimization trials suggests our method is better suited for applications where there is a strong prior over range and depth rather than for wide-area search. Cases where the approximate location of a source is known but precise localization is required could include localization of towed sources, underwater vehicles, and oceanographic moorings.

An important consideration for this method is that the quality of the optimization is contingent on the quality of the ambiguity surface. If the signal-to-noise ratio of the data is low, a sidelobe could be the global optimum rather than the peak corresponding to the true source location. Adequate signal processing steps must therefore be taken to maximize signal gain and reduce sidelobes. Approaches for reducing sidelobes employ high-resolution beamformers such

as SBL [25, 26, 27, 28] or multiple signal classification (MUSIC)[24] as the objective function. However, in optimization applications, conventional beamforming is often preferred as it has broad peaks and is more robust.

4.6 Conclusion

We have demonstrated efficient and accurate source localization with sequential Bayesian optimization when the ambiguity surface is modeled as a Gaussian process and sampling is guided by a probabilistic acquisition function. In addition to being sample-efficient, the Sobol+GP/EI and Sobol+GP/qEI strategies are advantageous as they are suitable for non-convex objective functions and require no information about the gradient of the objective function.

Simulations of a shallow-water waveguide and real data from an acoustic source tow experiment demonstrated that the Sobol+GP/EI and Sobol+GP/qEI strategies converge on the global optimum rapidly and yield superior results compared to grid search and quasi-random sampling strategies. We conclude the Sobol+GP/EI and Sobol+GP/qEI strategies are best employed as a hybrid sampling strategy in which the majority of trials in a fixed budget establish the prior over the objective, and the remaining Bayesian optimization trials fine-tune the optimization.

4.7 Acknowledgements

This research was supported by the Office of Naval Research through Grant No. N00014-21-1-2267 and the National Defense Science and Engineering Graduate Fellowship for Jenkins. Code implementation of this research uses the GPyTorch [62], BoTorch [15], and Adaptive Experimentation Platform [63] libraries. SBL was implemented with the SBL4 package [61].

Chapter 4, in full, is a reprint of the material as it appears in the Journal of the Acoustical Society of America 2023. Jenkins, William; Gerstoft, Peter; Park, Yongsung, Acoustical Society of America, 2023. The dissertation author was the primary investigator and author of this paper.

4.8 References

[1] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press, 2006.

[2] D. Caviedes-Nozal, N. A. B. Riis, F. M. Heuchel, J. Brunskog, P. Gerstoft, and E. Fernandez-Grande, "Gaussian processes for sound field reconstruction," *J. Acoust. Soc. Am.*, vol. 149, pp. 1107–1119, Feb. 2021.

[3] Z.-H. Michalopoulou, P. Gerstoft, and D. Caviedes-Nozal, "Matched field source localization with Gaussian processes," *JASA Express Lett.*, vol. 1, p. 064801, June 2021.

[4] Z.-H. Michalopoulou and P. Gerstoft, "Inversion in an uncertain ocean using Gaussian processes," *J. Acoust. Soc. Am.*, vol. 153, pp. 1600–1611, Mar. 2023.

[5] I. D. Khurjekar, P. Gerstoft, C. F. Mecklenbräuker, and Z.-H. Michalopoulou, "Direction-of-Arrival Estimation Using Gaussian Process Interpolation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 1–5, 2023.

[6] D. Zimmerman, C. Pavlik, A. Ruggles, and M. P. Armstrong, "An Experimental Comparison of Ordinary and Universal Kriging and Inverse Distance Weighting," *Math. Geol.*, vol. 31, pp. 375–390, May 1999.

[7] K. P. Murphy, *Probabilistic Machine Learning: An Introduction*. Cambridge, MA: MIT Press, 2022.

[8] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, "Taking the Human Out of the Loop: A Review of Bayesian Optimization," *Proc. IEEE*, vol. 104, pp. 148–175, Jan. 2016.

[9] J. J. Beland and P. B. Nair, "Bayesian Optimization Under Uncertainty," in *NIPS BayesOpt 2017 Workshop*, vol. 3, (Long Beach, Calif.), 2017.

[10] D. R. Jones, "A Taxonomy of Global Optimization Methods Based on Response Surfaces," *J. Global Optim.*, vol. 21, no. 4, pp. 345–383, 2001.

[11] N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger, "Information-Theoretic Regret Bounds for Gaussian Process Optimization in the Bandit Setting," *IEEE Trans. Inf. Theory*, vol. 58, no. 5, pp. 3250–3265, 2012.

[12] D. Ginsbourger, R. Le Riche, and L. Carraro, "Kriging Is Well-Suited to Parallelize Optimization," in *Computational Intelligence in Expensive Optimization Problems*, vol. 2, pp. 131–162, Berlin, Heidelberg: Springer, 2010.

[13] J. T. Wilson, F. Hutter, and M. P. Deisenroth, "Maximizing acquisition functions for Bayesian optimization," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[14] J. Wang, S. C. Clark, E. Liu, and P. I. Frazier, "Parallel Bayesian Global Optimization of Expensive Functions," *Oper. Res.*, vol. 68, pp. 1850–1865, Nov. 2020.

[15] M. Balandat, B. Karrer, D. R. Jiang, S. Daulton, B. Letham, A. G. Wilson, and E. Bakshy, "BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 21524–21538, 2020.

[16] A. Baggeroer, W. Kuperman, and P. Mikhalevsky, "An overview of matched field methods in ocean acoustics," *J. Ocean. Eng.*, vol. 18, pp. 401–424, Oct. 1993.

[17] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," *J. Mach. Learn. Res.*, vol. 13, no. 10, pp. 281–305, 2012.

[18] I. Sobol', "On the distribution of points in a cube and the approximate evaluation of integrals," *USSR Comp. Math. and Math. Phys.*, vol. 7, pp. 86–112, Jan. 1967.

[19] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola, *Global Sensitivity Analysis. The Primer*. West Sussex, U.K.: John Wiley & Sons, Ltd, 1 ed., Dec. 2007.

[20] S. Joe and F. Y. Kuo, "Constructing Sobol Sequences with Better Two-Dimensional Projections," *SIAM J. Sci. Comput.*, vol. 30, pp. 2635–2654, Jan. 2008.

[21] L. Dumaz, J. Garnier, and G. Lepoultier, "Acoustic and geoacoustic inverse problems in randomly perturbed shallow-water environments," *J. Acoust. Soc. Am.*, vol. 146, pp. 458–469, July 2019.

[22] B. Kayser, B. Gauvreau, D. Écotière, and V. Mallet, "Wind turbine noise uncertainty quantification for downwind conditions using metamodeling," *J. Acoust. Soc. Am.*, vol. 151, pp. 390–401, Jan. 2022.

[23] J. Nocedal and S. J. Wright, *Numerical Optimization*. Springer Series in Operations Research, New York: Springer, 2nd ed., 2006.

[24] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, pp. 276–280, Mar. 1986.

[25] M. E. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.

[26] D. Wipf and B. Rao, "Sparse Bayesian Learning for Basis Selection," *IEEE Trans. Signal Process.*, vol. 52, pp. 2153–2164, Aug. 2004.

[27] K. L. Gemba, S. Nannuru, P. Gerstoft, and W. S. Hodgkiss, "Multi-frequency sparse Bayesian learning for robust matched field processing," *J. Acoust. Soc. Am.*, vol. 141, pp. 3411–3420, May 2017.

[28] Y. Park, F. Meyer, and P. Gerstoft, "Sequential sparse Bayesian learning for time-varying direction of arrival," *J. Acoust. Soc. Am.*, vol. 149, pp. 2089–2099, Mar. 2021.

[29] Y. Chi, Y. M. Lu, and Y. Chen, "Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview," *IEEE Trans. Signal Process.*, vol. 67, pp. 5239–5269, Oct. 2019.

[30] S. Li, L. Cheng, T. Zhang, H. Zhao, and J. Li, "Graph-guided Bayesian matrix completion for ocean sound speed field reconstruction," *The Journal of the Acoustical Society of America*, vol. 153, pp. 689–710, Jan. 2023.

[31] M. J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M. A. Roch, S. Gannot, and C.-A. Deledalle, "Machine learning in acoustics: Theory and applications," *The Journal of the Acoustical Society of America*, vol. 146, pp. 3590–3628, Nov. 2019.

[32] Y. Liu, H. Niu, and Z. Li, "A multi-task learning convolutional neural network for source localization in deep ocean," *J. Acoust. Soc. Am.*, vol. 148, pp. 873–883, Aug. 2020.

[33] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *J. Acoust. Soc. Am.*, vol. 152, pp. 107–151, July 2022.

[34] A. Varon, J. Mars, and J. Bonnel, "Approximation of modal wavenumbers and group speeds in an oceanic waveguide using a neural network," *JASA Express Lett.*, vol. 3, p. 066003, June 2023.

[35] L. Zhang, T. Zhang, H.-S. Shin, and X. Xu, "Efficient Underwater Acoustical Localization Method Based On Time Difference and Bearing Measurements," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–16, 2021.

[36] P. Gerstoft, "Inversion of seismoacoustic data using genetic algorithms and a posteriori probability distributions," *J. Acoust. Soc. Am.*, vol. 95, pp. 770–782, Feb. 1994.

[37] P. Gerstoft and C. F. Mecklenbräuker, "Ocean acoustic inversion with estimation of *a posteriori* probability distributions," *J. Acoust. Soc. Am.*, vol. 104, pp. 808–819, Aug. 1998.

[38] S. E. Dosso, "Quantifying uncertainty in geoacoustic inversion. I. A fast Gibbs sampler approach," *J. Acoust. Soc. Am.*, vol. 111, pp. 129–142, Jan. 2002.

[39] S. E. Dosso and P. L. Nielsen, "Quantifying uncertainty in geoacoustic inversion. II. Application to broadband, shallow-water data," *J. Acoust. Soc. Am.*, vol. 111, pp. 143–159, Jan. 2002.

[40] S. E. Dosso and M. J. Wilmut, "Uncertainty estimation in simultaneous Bayesian tracking and environmental inversion," *J. Acoust. Soc. Am.*, vol. 124, pp. 82–97, July 2008.

[41] M. D. Collins, W. A. Kuperman, and H. Schmidt, "Nonlinear inversion for ocean-bottom properties," *J. Acoust. Soc. Am.*, vol. 92, pp. 2770–2783, Nov. 1992.

[42] M. D. Collins, WA. Kuperman, and W. L. Siegmann, "Propagation and inversion in complex ocean environments," in *Full Field Inversion Methods in Ocean and Seismo-Acoustics*, pp. 15–20, The Netherlands: Kluwer Academic Publishers, 1995.

[43] M. D. Collins and L. Fishman, "Efficient navigation of parameter landscapes," *J. Acoust. Soc. Am.*, vol. 98, no. 3, pp. 1637–1644, 1995.

[44] J. Dettmer, S. E. Dosso, and C. W. Holland, "Trans-dimensional geoacoustic inversion," *J. Acoust. Soc. Am.*, vol. 128, pp. 3393–3405, Dec. 2010.

[45] J. Dettmer and S. E. Dosso, "Trans-dimensional matched-field geoacoustic inversion with hierarchical error models and interacting Markov chains," *J. Acoust. Soc. Am.*, vol. 132, pp. 2239–2250, Oct. 2012.

[46] S. E. Dosso and J. Bonnel, "Joint trans-dimensional inversion for water-column sound speed and seabed geoacoustic models," *JASA Express Lett.*, vol. 3, p. 060801, June 2023.

[47] P. Gerstoft, "Inversion of acoustic data using a combination of genetic algorithms and the Gauss–Newton approach," *J. Acoust. Soc. Am.*, vol. 97, pp. 2181–2190, Apr. 1995.

[48] M. R. Fallat and S. E. Dosso, "Geoacoustic inversion via local, global, and hybrid algorithms," *J. Acoust. Soc. Am.*, vol. 105, pp. 3219–3230, June 1999.

[49] N. Booth, P. Baxley, J. Rice, P. Schey, W. Hodgkiss, G. D'Spain, and J. Murray, "Source localization with broad-band matched-field processing in shallow water," *J. Ocean. Eng.*, vol. 21, no. 4, pp. 402–412, 1996.

[50] Z.-H. Michalopoulou and M. Porter, "Matched-field processing for broad-band source localization," *J. Ocean. Eng.*, vol. 21, no. 4, pp. 384–392, 1996.

[51] R. M. Neal, *Bayesian Learning for Neural Networks*, vol. 118 of *Lecture Notes in Statistics*. New York: Springer, 1996.

[52] K. P. Murphy, *Probabilistic Machine Learning: Advanced Topics*. Cambridge, MA: MIT Press, 2023.

[53] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A Limited Memory Algorithm for Bound Constrained Optimization," *SIAM J. Sci. Comput.*, vol. 16, no. 5, pp. 1190–1208, 1995.

[54] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimization," *ACM Trans. Math. Software*, vol. 23, pp. 550–560, Dec. 1997.

[55] S. Kim, G. F. Edelmann, W. A. Kuperman, W. S. Hodgkiss, H. C. Song, and T. Akal, "Spatial resolution of time-reversal arrays in shallow water," *J. Acoust. Soc. Am.*, vol. 110, pp. 820–829, Aug. 2001.

[56] B. Letham, B. Karrer, G. Ottoni, and E. Bakshy, "Constrained Bayesian Optimization with Noisy Experiments," June 2018.

[57] R. T. Bachman, P. W. Schey, N. O. Booth, and F. J. Ryan, "Geoacoustic databases for matched-field processing: Preliminary results in shallow water off San Diego, California," *J. Acoust. Soc. Am.*, vol. 99, pp. 2077–2085, Apr. 1996.

[58] Marine Physical Laboratory, "SWellEx-96 Experiment." http://swellex96.ucsd.edu/.

[59] M. B. Porter, "The KRAKEN normal mode program," *SACLANT Undersea Research Centre Memorandum (SM-245)/Naval Research Laboratory Memorandum Report 6920*, Sept. 1991.

[60] G. L. D'Spain, J. J. Murray, and W. S. Hodgkiss, "Mirages in shallow water matched field processing," *J. Acoust. Soc. Am.*, vol. 105, pp. 3245–3265, June 1999.

[61] Y. Park, S. Nannuru, K. Gemba, and P. Gerstoft, "SBL4 from NoiseLab." https://github.com/gerstoft/SBL, 2020.

[62] J. R. Gardner, G. Pleiss, D. Bindel, K. Q. Weinberger, and A. G. Wilson, "GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[63] Meta Platforms, Inc., "Adaptive Experimentation Platform." https://ax.dev, 2023.

Chapter 5

Bayesian optimization with Gaussian processes for robust localization

We present a sample-efficient Bayesian optimization (BO) method to estimate underwater source localization robust to unknown tilt in a vertical line array. Rather than conducting exhaustive search of parameter space to estimate localization and tilt, BO uses a Gaussian process (GP) surrogate model of the Bartlett power objective function to guide sampling of the parameter space. Samples are suggested using a heuristic acquisition function that uses the GP to balance exploitation and exploration of parameter space. Using experimental data, we show that BO obtains better localization estimates than conventional grid search and quasi-random sampling strategies, and that robustness to array tilt comes with little additional computational cost.

5.1 Introduction

Localization of an underwater acoustic source with a vertical line array (VLA) can be accomplished through matched field processing (MFP) [1], which seeks to match replica pressure fields produced by a propagation model to the observed field on the VLA. MFP can be computationally intensive, requiring many evaluations of a propagation model evaluated in a grid search of the parameter space. The computational cost of grid search is exacerbated when the dimensionality of the parameter space increases; compensatory measures like reducing

grid resolution risk missing the optimal matched field and may lead to localization errors. Because MFP is sensitive to array tilt, which arises from currents acting on the VLA [2], [3], [4], estimating tilt to calibrate the VLA introduces an additional parameter to the search space and increases computational cost.

Source localization robust to array tilt has been demonstrated using sparse Bayesian learning (SBL) [5], [6] and multiple constraint MFP (MCM) [4]. However, both SBL and MCM are implemented as grid search and can be computationally expensive. Markov chain Monte Carlo methods [7], [8] and particle filtering [9] have also been used to estimate source localization and array tilt, among other geoacoustic parameters, but rely on thousands of evaluations of the forward model and require careful hyperparameter tuning. Data-driven localization methods using deep learning and neural networks have been demonstrated, but require large amounts of training data specific to a particular scenario [10], [11].

Bayesian optimization (BO) is a sample-efficient sequential optimization method for estimating the global optimum of an objective function [12], [13] and has been implemented for a variety of signal processing tasks [14], [15], [16], [17]. The framework is sequential in that it suggests a new sample in parameter space to evaluate based on previous evaluations of the objective function. Performance and characterization of BO were evaluated for underwater source localization with a known tilt [18].

We propose utilizing the sample-efficient BO framework to estimate source localization robust to array tilt mismatch by including tilt as a search parameter. Given a fixed amount of parameter space samples, we demonstrate that BO outperforms conventional methods like grid search and quasi-random search in estimating source localization. Furthermore, we demonstrate that BO obtains accurate estimates of source localization more rapidly than conventional methods.

5.2 Parameterization

We define our model and objective function after [7]. Consider an acoustic pressure field $\mathbf{q}(\omega_l) \in \mathbb{C}^J$ sampled by a vertical line array (VLA) with J elements processed at frequencies $\mathbf{\Omega} = [\omega_1, \ldots, \omega_L]$. Observed data are the sum of predictions of the data $\mathbf{p}(\mathbf{m}, \omega_l) \in \mathbb{C}^J$ given a model $\mathbf{m} \in \mathscr{M}^D$ with D parameters, and an error term $\mathbf{e}(\omega_l)$:

$$\mathbf{q}(\omega_l) = \mathbf{p}(\mathbf{m}, \omega_l) + \mathbf{e}(\omega_l). \tag{5.1}$$

For joint localization and array tilt estimation, the model is parameterized as $\mathbf{m} = [r_s, z_s, \tau]^\mathsf{T}$, where r_s is source range, z_s is source depth, and τ is array tilt. Predicted data are modeled by:

$$\mathbf{p}(\mathbf{m}, \omega_l) = \mathbf{w}(\mathbf{m}, \omega_l) S(\omega_l) \tag{5.2}$$

where $\mathbf{w}(\mathbf{m}, \omega_l) \in \mathbb{C}^J$ is a replica pressure field produced by an acoustic propagation model, and $S(\omega_l) \in \mathbb{C}$ is an unknown deterministic source term. For brevity, we denote frequency dependence by $\mathbf{q}_l = \mathbf{q}(\omega_l)$, etc.

Given VLA element depths $\mathbf{z} = [z_1, \ldots, z_J]^\mathsf{T}$, transformed coordinates $\mathbf{r}'$ and $\mathbf{z}'$ account for a vertical line array (VLA) anchored at depth z_b and range r_s, tilted by τ from vertical in the source-receiver plane:

$$\mathbf{r}' = r_s + (z_b - \mathbf{z}) \sin \tau \tag{5.3}$$

$$\mathbf{z}' = (z_b - \mathbf{z}) \cos \tau. \tag{5.4}$$

The replica pressure field is parameterized as a summation of modes indexed by m:

$$\mathbf{w}_l(r_s, z_s, \tau) = \frac{1}{4\rho(z_s)} \sum_{m=1}^{\infty} \Psi_m(z_s) \Psi_m(\mathbf{z}') H_0^{(1)}\left(k_{rm} \mathbf{r}'\right) \tag{5.5}$$

where ρ is water density; Ψ_m is the mth mode function at a given depth; $H_0^{(1)}$ is the zeroth-order

Hankel function of the first kind; and k_{rm} is the horizontal wavenumber of the mth mode [19]. Ψ and k_{rm} are computed using KRAKEN [20].

The estimated model $\widehat{\mathbf{m}}$ is obtained by matching $\mathbf{q}_l$ and $\mathbf{p}_l$ through minimization of an objective function ϕ:

$$\widehat{\mathbf{m}} = \arg\min_{\mathscr{M}} [\phi(\mathbf{m})] \tag{5.6}$$

ϕ is constructed from the Bartlett objective function [7]:

$$\phi(\mathbf{m}) = \prod_{l=1}^{L} \left(\text{tr}\, \widehat{\mathbf{R}}_l - \frac{\mathbf{w}_l^{\mathsf{H}}(\mathbf{m}) \widehat{\mathbf{R}}_l \mathbf{w}_l(\mathbf{m})}{\mathbf{w}_l^{\mathsf{H}}(\mathbf{m}) \mathbf{w}_l(\mathbf{m})} \right) \tag{5.7}$$

where $\widehat{\mathbf{R}}_l \in \mathbb{C}^{J \times J}$ is the sample covariance matrix (SCM) from K snapshots:

$$\widehat{\mathbf{R}}_l = \frac{1}{K} \sum_{k=1}^{K} \mathbf{q}_{k,l} \mathbf{q}_{k,l}^{\mathsf{H}}. \tag{5.8}$$

5.3 Bayesian Optimization Framework

In BO, a GP surrogate model $\mathscr{GP}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (defined below) approximates the objective function $\phi(\mathbf{m})$ by performing GP regression over parameter space $\mathscr{M}$. The GP surrogate model is then used by an acquisition function to suggest a new sample in $\mathscr{M}$ to evaluate. The process repeats as data are added by exploration of $\mathscr{M}$ until a fixed budget of N_{total} evaluations, or trials, is expended.

5.3.1 Gaussian process regression

Here we follow the derivations of [21]. A Gaussian process is a collection of N random points:

$$\mathbf{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_N] \in \mathscr{M}^{D \times N}, \tag{5.9}$$

any finite of number of which have a joint Gaussian distribution. Given a real process $\mathbf{f} = [\phi(\mathbf{m}_1), \ldots, \phi(\mathbf{m}_N)]^\mathsf{T}$, a GP is described by a mean function,

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{f}] = [\mu(\mathbf{m}_1), \ldots, \mu(\mathbf{m}_N)]^\mathsf{T} \in \mathbb{R}^N, \tag{5.10}$$

where $\mu(\mathbf{m}_n)$ is the mean at $\mathbf{m}_n$; and a covariance function,

$$\boldsymbol{\Sigma}_{ij} = \mathbb{E}[(f(\mathbf{m}_i) - \mu(\mathbf{m}_i))(f(\mathbf{m}_j) - \mu(\mathbf{m}_j))] \tag{5.11}$$

$$= \mathscr{K}(\mathbf{m}_i, \mathbf{m}_j) \in \mathbb{R}^{N \times N}, \tag{5.12}$$

where $\mathscr{K}(\mathbf{m}_i, \mathbf{m}_j)$ is a kernel function measuring the similarity between points $\mathbf{m}_i$ and $\mathbf{m}_j$. The covariance function is assumed diagonal such that the variance at a point $\sigma(\mathbf{m})$ is obtained from $\boldsymbol{\Sigma} \approx \text{diag}\left([\sigma^2(\mathbf{m}_1), \ldots, \sigma^2(\mathbf{m}_N)]\right)$. The real process $\mathbf{f}$ is modeled by the GP as $\mathbf{f} \sim \mathscr{GP}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and has observations comprising parameters evaluated by the objective function:

$$\mathscr{D} = \{(\mathbf{m}_n, y_n),\ n = 1 : N\} = \{\mathbf{M}, \mathbf{y}\}, \quad \mathbf{y} \in \mathbb{R}^N. \tag{5.13}$$

The GP predicts N^* unobserved outputs $\mathbf{f}_*$ at inputs $\mathbf{M}_{*,D\times N^*} = [\mathbf{m}_1^*, \ldots, \mathbf{m}_{N^*}^*]$. From [7, eq. (17.33)], the joint distribution of the observed process $\mathbf{y}$ and the predictive distribution $\mathbf{f}_*$ is:

$$p(\mathbf{y}, \mathbf{f}_* | \mathbf{M}, \mathbf{M}_*) = \begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} = \mathscr{N}\left(\begin{bmatrix} \boldsymbol{\mu}_M \\ \boldsymbol{\mu}_* \end{bmatrix}, \begin{bmatrix} \widehat{\mathbf{K}}_{M,M} & \mathbf{K}_{M,*} \\ \mathbf{K}_{M,*}^\mathsf{T} & \mathbf{K}_{*,*} \end{bmatrix} \right) \tag{5.14}$$

where $\boldsymbol{\mu}_M$ and $\boldsymbol{\mu}_*$ are the mean functions at $\mathbf{M}$ and $\mathbf{M}_*$, respectively; and

$$\widehat{\mathbf{K}}_{M,M} = \mathbf{K}_{M,M} + \sigma_y^2 \mathbf{I} = \mathscr{K}(\mathbf{M},\mathbf{M})^{N\times N} + \sigma_y^2 \mathbf{I} \tag{5.15}$$

$$\mathbf{K}_{M,*} = \mathscr{K}(\mathbf{M},\mathbf{M}_*)^{N\times N_*} \tag{5.16}$$

$$\mathbf{K}_{*,*} = \mathscr{K}(\mathbf{M}_*,\mathbf{M}_*)^{N_*\times N_*}. \tag{5.17}$$

From [7, eq. (17.34)], the GP is conditioned on the new observations, giving the posterior distribution:

$$p(\mathbf{f}_*|\mathscr{D},\mathbf{M}_*) = \mathscr{N}(\mathbf{f}_*|\boldsymbol{\mu}_{*|M},\boldsymbol{\Sigma}_{*|M}) \tag{5.18}$$

$$\boldsymbol{\mu}_{*|M} = \boldsymbol{\mu}_* + \mathbf{K}_{M,*}^{\mathsf{T}} \widehat{\mathbf{K}}_{M,M}^{-1}(\mathbf{y} - \boldsymbol{\mu}_M) \tag{5.19}$$

$$\boldsymbol{\Sigma}_{*|M} = \mathbf{K}_{*,*} - \mathbf{K}_{M,*}^{T} \widehat{\mathbf{K}}_{M,M}^{-1} \mathbf{K}_{M,*}. \tag{5.20}$$

The kernel function (5.12) measures similarity between two points in $\mathscr{M}$. We use a positive-definite and stationary kernel with real-valued inputs, i.e.,

$$\mathscr{K}(\mathbf{m}_i,\mathbf{m}_j) = \mathscr{K}(\mathbf{r}), \quad \mathbf{r} = \mathbf{m}_i - \mathbf{m}_j \tag{5.21}$$

and specifically adopt the Matern kernel with smoothness parameter $\nu = 5/2$:

$$\mathscr{K}\left(\mathbf{r};\frac{5}{2},\mathbf{l}\right) = \sigma_y^2 \prod_{d=1}^{D} \left(1 + \frac{\sqrt{5}r_d}{l_d} + \frac{5r_d^2}{3l_d^2}\right) \exp\left(-\frac{\sqrt{5}r_d}{l_d}\right) \tag{5.22}$$

where $\sigma_y \in \mathbb{R}$ and $l_d \in \mathbb{R}$ are hyperparameters, with σ_y^2 estimating the noise variance of the GP and $\mathbf{l} = [l_1,\ldots,l_D]$ controlling the length scale in dimension d [7, eq. (17.13)].

Kernel hyperparameters $\boldsymbol{\theta} = [\sigma_y^2,\mathbf{l}]$ must be optimized for the GP surrogate model to appropriately reflect the data and is efficiently performed with an empirical-Bayes approach.

Noting $\boldsymbol{\theta}$ is implicit in $\widehat{\mathbf{K}}_{M,M}$, marginalizing the product of the Gaussians

$$p(\mathbf{f}|\mathbf{M},\boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}_M, \widehat{\mathbf{K}}_{M,M}) \tag{5.23}$$

$$p(\mathbf{y}|\mathbf{f},\mathbf{M}) = \prod_{n=1}^{N} \mathcal{N}(y_n|f_n,\sigma_y^2). \tag{5.24}$$

with respect to $\mathbf{f}$ results in the marginal likelihood, itself a Gaussian [7, eq. (17.51)]:

$$p(\mathbf{y}|\mathbf{M},\boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f},\mathbf{M},\boldsymbol{\theta}) p(\mathbf{f}|\mathbf{M},\boldsymbol{\theta}) d\mathbf{f} \tag{5.25}$$

$$= \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_M, \widehat{\mathbf{K}}_{M,M}). \tag{5.26}$$

To find the optimal $\boldsymbol{\theta}$, the log marginal likelihood is maximized [7, eq. (17.51)]:

$$\begin{aligned} L &= \log p(\mathbf{y}|\mathbf{M},\boldsymbol{\theta}) = \log \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_M, \widehat{\mathbf{K}}_{M,M}) \\ &= -\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_M)^{\mathsf{T}} \widehat{\mathbf{K}}_{M,M}^{-1}(\mathbf{y}-\boldsymbol{\mu}_M) \\ &\quad -\frac{1}{2}\log|\widehat{\mathbf{K}}_{M,M}| - \frac{N}{2}\log(2\pi) \end{aligned} \tag{5.27}$$

$$\frac{\partial L}{\partial \theta_j} = \frac{1}{2}\mathrm{tr}\left[(\boldsymbol{\gamma}\boldsymbol{\gamma}^{\mathsf{T}} - \widehat{\mathbf{K}}_{M,M})\frac{\partial \widehat{\mathbf{K}}_{M,M}}{\partial \theta_j}\right], \tag{5.28}$$

where $\boldsymbol{\gamma} = \widehat{\mathbf{K}}_{M,M}^{-1}(\mathbf{y}-\boldsymbol{\mu}_M)$ [7, eq. (17.52)]. Since (5.27) is smoothly varying with few optima, hyperparameter optimization is performed with the bounded Limited-memory BFGS (L-BFGS-B) algorithm [22], [23].

5.3.2 Acquisition function

At trial t, a sample $\mathbf{m}_t$ is evaluated by the objective function $\phi(\mathbf{m}_t)$ (5.7). The next sample $\mathbf{m}_{t+1}$ is suggested through optimization of a heuristic acquisition function $\alpha(\phi(\mathbf{m}))$:

$$\mathbf{m}_{t+1} = \underset{\mathbf{m}\in\mathscr{M}}{\arg\max}\left[\alpha(\phi(\mathbf{m}))\right] \tag{5.29}$$

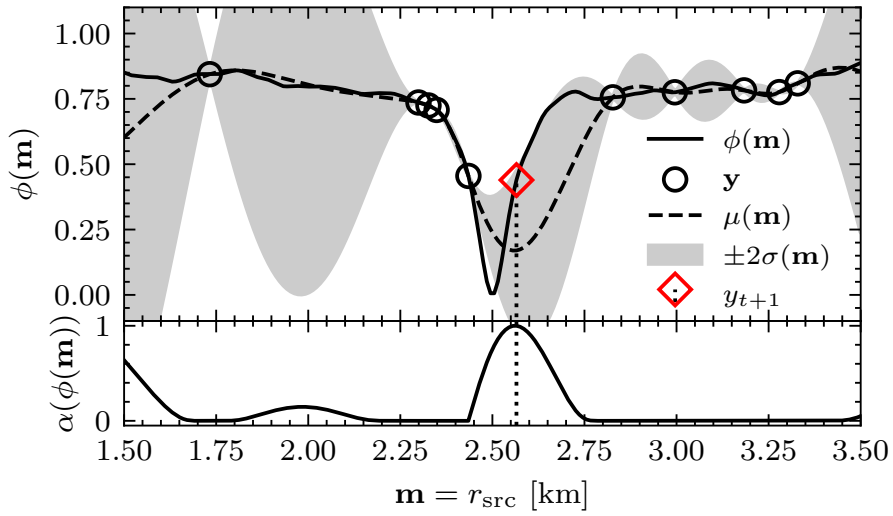

Figure 5.1. Gaussian process regression (upper panel) of the objective function $\phi(\mathbf{m})$ for one-dimensional ambiguity surface over source range. The true surface (solid) is approximated by the mean function $\mu(\mathbf{m})$ (dashed) and uncertainty $\sigma(\mathbf{m})$ (shaded) conditioned on observed data $\mathbf{y}$ (dots). The next sample y_{t+1} is suggested by the maximum of the acquisition function $\alpha(\mathbf{m})$ (lower panel) normalized to $[0, 1]$.

Table 5.1. Pseudocode for Bayesian optimization with GP surrogate model.

Input: Parameter domain $\mathscr{M}$, objective function ϕ, kernel function $\mathscr{K}$, acquisition function α, total trials N_{total}.
Output: Best estimate of parameters $\widehat{\mathbf{m}}$

$\widehat{\mathbf{m}}, \phi' \leftarrow \textsc{Sobol}[\mathscr{M}, N_{\mathrm{init}}]$
for $t = N_{\mathrm{init}} + 1$ **to** N_{total} **do**
  $\boldsymbol{\mu} \leftarrow \mathbb{E}[\phi(\mathbf{M})]$ [eq. (5.10)]
  $\boldsymbol{\theta} \leftarrow \text{L-BFGS-B}\,[\mathscr{K}(\mathbf{M}, \mathbf{M}; \boldsymbol{\theta})]$ [eq. (5.27)-(5.28)]
  $\boldsymbol{\Sigma} \leftarrow \mathscr{K}(\mathbf{M}, \mathbf{M}; \boldsymbol{\theta})$ [eq. (5.12)]
  $\mathbf{m}_t \leftarrow \arg\max_{\mathbf{m}} \alpha(\phi(\mathbf{m}))$ [Section 5.3.2]
  $\phi_t \leftarrow \phi(\mathbf{m}_t)$
  if $\phi_t > \phi'$ **then**
    $\widehat{\mathbf{m}} \leftarrow \mathbf{m}_t$, $\phi' \leftarrow \phi_t$

Table 5.2. Bayesian optimization implementation parameters.

Parameter	Description	Value
N_{total}	Total trials	64
N_{init}	Warm-up trials	32
N_{acq}	Samples for acquisition function optimization	1024
N_{restart}	Acquisition function re-starts	40

An ideal acquisition function balances exploration of regions of high uncertainty with exploitation (i.e., dense sampling) of well-performing regions. We adopt the expected improvement (EI) acquisition function due to its ability to adaptively balance these competing goals [24], [13]. EI quantifies the amount of improvement a sample is expected to yield over the lowest previously observed objective function ϕ', and is defined as:

$$\begin{aligned} \alpha_{\mathrm{EI}}(\phi(\mathbf{m})) &= \mathbb{E}\left[\max(\phi(\mathbf{m}) - \phi', 0)\right] \\ &= \sigma(\mathbf{m})(z\Phi(z) + \varphi(z)) \end{aligned} \tag{5.30}$$

where z is the standardized improvement:

$$z = \frac{\phi' - \mu(\mathbf{m})}{\sigma(\mathbf{m})}, \tag{5.31}$$

Φ is the cumulative distribution function, and φ is the normal probability distribution function.

5.3.3 Implementation

BO is implemented using the BoTorch Python library [25], with pseudocode given in Table 5.1 and algorithm parameters in Table 5.2. To initialize the GP surrogate model, N_{init} warmup trials are completed by using a Sobol sequence [26] to generate quasi-random samples from the parameter space $\mathscr{M}$. BO proceeds by performing the GP regression on the observed data $\mathscr{D}$ and optimizing the acquisition function α to suggest the next sample. The process repeats until a fixed budget of N trials is expended. Since the acquisition function can be non-convex but is inexpensive to evaluate, (5.30) is optimized by taking N_{restart} random restarts of L-BFGS-B initialized with N_{acq} samples drawn from the acquisition function; values for N_{restart} and N_{init} are given in Table 5.2.

One iteration of BO for a one-dimensional ambiguity surface over source range is shown in Fig. 5.1 with the GP surrogate model for the objective function $\phi(\mathbf{m})$ and the corresponding acquisition function $\alpha(\phi(\mathbf{m}))$. In this example, the GP surrogate model was fit to 10 samples

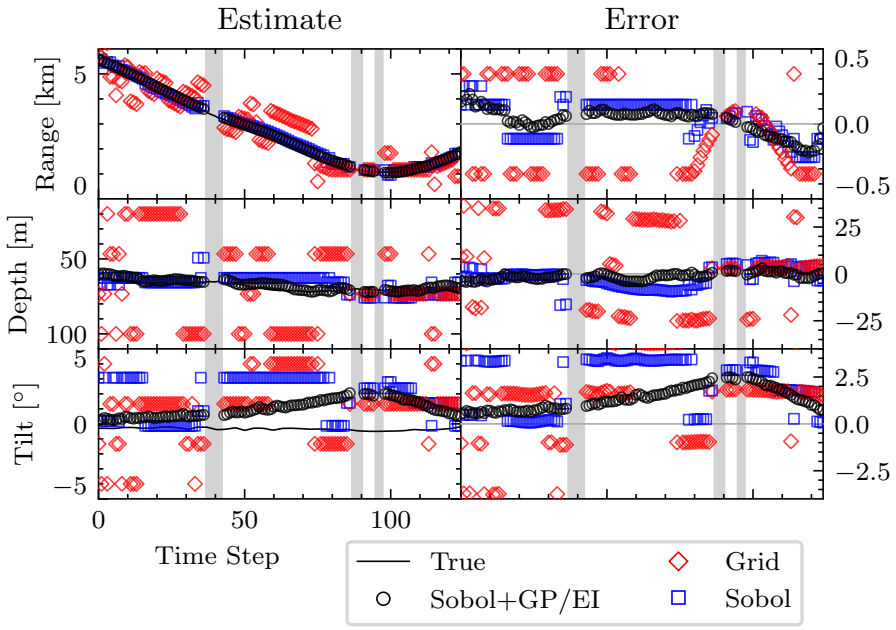

Figure 5.2. Parameter estimates and errors. Gray regions indicate when the source ceased transmitting.

from the objective function. The EI acquisition function is non-zero in regions of high uncertainty and low objective function value, and zero in regions of low uncertainty and high objective function value. The next sample is suggested by the maximum of the acquisition function.

5.4 Experimental Results

Localization and array tilt estimation are demonstrated using experimental data collected during SWellEx-96 off the coast of Southern California [27]. A 64-element VLA was anchored at the ocean bottom at 217 m depth, with elements spaced evenly between 94.125 m and 212.25 m (1.875 m spacing). During event S5, an acoustic source was towed in a straight line from south to north at 5 knots, during which time the source closed from 5 km to within 1 km at the closest point of approach (CPA) before opening to 2 km. At CPA, the VLA stood nearly due west of the source. The source was towed at 60 m depth and transmitted 13 tones between 50 and 400 Hz.

Data were recorded at the VLA with a 1.5 kHz sampling rate and processed in 8,192-sample (2.7 s) segments with 50% overlap. A Hann window was applied to each segment and

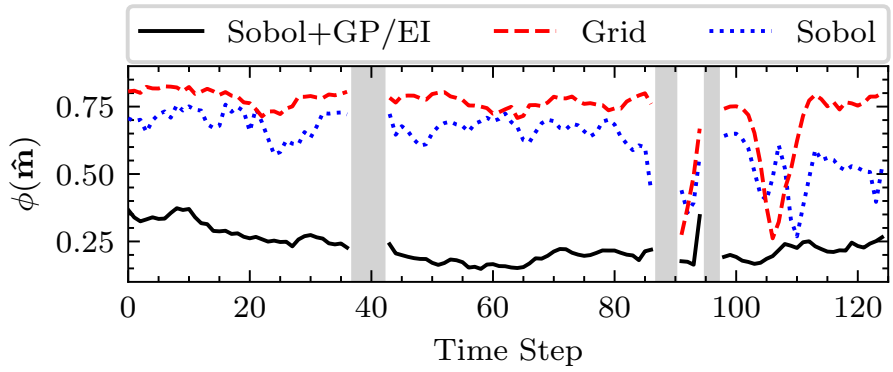

Figure 5.3. Lowest observed objective function $\phi(\widehat{\mathbf{m}})$ over 64 trials for each optimization strategy.

Table 5.3. Mean absolute error (MAE) of strategies over all time steps.

Strategy	r_{src} [km]	z_{src} [m]	τ [°]
Grid	0.49	23.7	2.2
Sobol	0.12	4.7	2.2
Sobol+GP/EI	**0.08**	**1.8**	**1.3**

8,192-point FFT computed, from which complex acoustic pressure on the array was retrieved for the 13 source frequencies. SCMs were computed (5.8) with $K = 8$ overlapping segments. Processing resulted in discretization of 45 minutes of data into 125 time steps. On three occasions, the source ceased transmitting; affected segments were removed.

Forward model computations for (5.7) were performed using an acoustic environment model [27], [28]. Of note, though the model is range-independent, bottom depth at the source was up to 30 m shallower than at the VLA; true source ranges and depths were corrected according to [29]. The parameter search space was defined as $\widehat{r}_s \in r_s \pm 1.0$ km, $\widehat{z}_s \in z_s \pm 40$ m, and $\widehat{\tau} \in [-4°, 4°]$. True source range r_s was obtained from a GPS receiver on the tow vessel, and source depth z_s was set to 60 m; both values were corrected according to [29]. Tilt data and array heading were measured by an inclinometer on the VLA and used to calculate τ.

Three estimation strategies were evaluated for each time step: grid search, quasi-random search using Sobol sequences [26], and BO. BO was performed using a strategy of warmup trials generated by a Sobol sequence followed by BO trials, i.e., the Sobol+GP/EI strategy. From

Table 5.2, 64 total trials were completed, with 32 warmup trials and 32 BO trials. Optimization was re-initialized at every time step with only search space constraints as prior information. To compare performance for an equivalent number of trials, Sobol sampling and grid search were also evaluated with 64 trials, with grid search parameter space discretized to a $4 \times 4 \times 4$ grid.

Fig. 5.2 shows parameter estimates and associated errors for source range, depth, and VLA tilt at each time step for the 64-trial optimizations. BO (Sobol + GP/EI) consistently estimates source range and depth with minimal error. Array tilt is estimated with a bias from the expected values but is consistent with previous MFP results [27], [6]; discrepancy in the estimated tilt is likely due to mismatch in the range-independent model environment. Grid search results illustrate an inadequately discretized parameter space, with low error occurring only when the sample point coincides with the optimum value of the objective function. Sobol sampling has better performance than grid search for the equivalent number of trials, but the parameter space remains inadequately sampled. See Table 5.3 for mean absolute errors of the estimated parameters over the source tow. Fig. 5.3 shows the lowest values of (5.7) for each time step, with BO out-performing grid and Sobol search.

In practice, grid search and quasi-random search for source localization and tilt estimation would be conducted with more than 64 trials with additional computational cost. To illustrate that BO is more time-efficient than grid search and Sobol sampling, Fig. 5.4 shows the progression of the lowest value of (5.7) vs. run time for all time steps but with greater sampling density for grid search and Sobol sampling. Grid search was discretized to $r_{\text{src}} \times z_{\text{src}} \times \tau \sim 24 \times 9 \times 5$ (1,080 trials), and Sobol sampling was performed with 1,024 trials. Even with the higher sampling density, grid search and Sobol sampling do not reach the low values of BO and would require increased sampling density to achieve equivalent performance.

Fig. 5.5 shows the lowest observed values of (5.7) for 64 trials of BO including estimated array tilt and BO with array tilt fixed to 0°. For little additional computational cost, including tilt in the parameter search space results in lower values of (5.7) than fixing tilt to a constant value. Furthermore, the robustness to array tilt is achieved with less computational cost than that

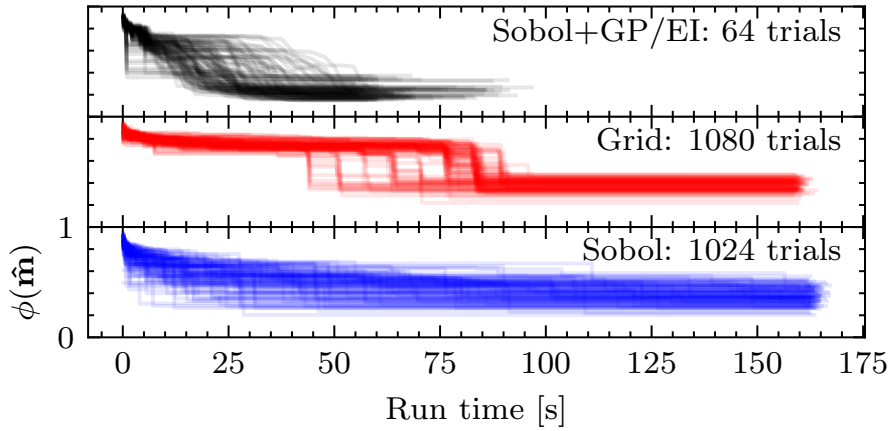

Figure 5.4. Lowest observed objective function $\phi(\hat{\mathbf{m}})$ vs. run time traced for all time steps.

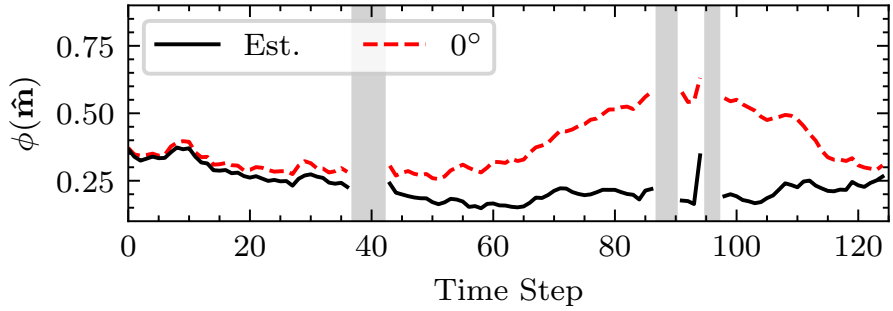

Figure 5.5. Lowest observed objective function $\phi(\hat{\mathbf{m}})$ for 64 trials of BO with tilt included in the parameter space (solid) and no tilt (dashed).

required by grid search and quasi-random sampling to achieve similar results (Fig. 5.4).

5.5 Conclusion

BO with a GP surrogate model is a sample-efficient strategy for estimating source localization in the presence of unknown array tilt. By leveraging the efficiency of BO, the inclusion of tilt in the parameter search space is computationally straightforward and leads to more robust estimates of source range and depth than when tilt is fixed. Furthermore, BO is able to achieve better performance than grid search and Sobol sampling in fewer trials and in less time.

5.6 Acknowledgements

Chapter 5, in full, has been submitted for publication of the material as it may appear in the Proceedings of the Institute of Electrical and Electronics Engineers International Conference on Acoustics, Speech, and Signal Processing 2024. Jenkins, William; Gerstoft, Peter, Institute of Electrical and Electronics Engineers, 2023. The dissertation author was the primary investigator and author of this paper.

5.7 References

[1] A. B. Baggeroer, W. A. Kuperman, and P. N. Mikhalevsky, "An overview of matched field methods in ocean acoustics," *IEEE J. Ocean. Eng.*, vol. 18, pp. 401–424, Oct. 1993.

[2] W. S. Hodgkiss, D. E. Ensberg, J. J. Murray, G. L. D'Spain, N. O. Booth, and P. W. Schey, "Direct measurement and matched-field inversion approaches to array shape estimation," *IEEE J. Oceanic Eng.*, vol. 21, pp. 393–401, Oct. 1996.

[3] G. Byun, C. Cho, H. C. Song, J. S. Kim, and S.-H. Byun, "Array invariant-based calibration of array tilt using a source of opportunity," *J. Acoust. Soc. Am.*, vol. 143, pp. 1318–1325, Mar. 2018.

[4] G. Byun, F. H. Akins, K. L. Gemba, H. C. Song, and W. A. Kuperman, "Multiple constraint matched field processing tolerant to array tilt mismatch," *J. Acoust. Soc. Am.*, vol. 147, pp. 1231–1238, Feb. 2020.

[5] K. L. Gemba, S. Nannuru, P. Gerstoft, and W. S. Hodgkiss, "Multi-frequency sparse Bayesian learning for robust matched field processing," *J. Acoust. Soc. Am.*, vol. 141, pp. 3411–3420, May 2017.

[6] K. L. Gemba, S. Nannuru, and P. Gerstoft, "Robust Ocean Acoustic Localization With Sparse Bayesian Learning," *IEEE J. Sel. Top. Signal Process.*, vol. 13, pp. 49–60, Mar. 2019.

[7] P. Gerstoft and C. F. Mecklenbräuker, "Ocean acoustic inversion with estimation of *a posteriori* probability distributions," *J. Acoust. Soc. Am.*, vol. 104, pp. 808–819, Aug. 1998.

[8] J. Bonnel, S. E. Dosso, J. A. Goff, Y. Lin, J. H. Miller, G. Potty, P. Wilson, and D. Knobles, "Transdimensional Geoacoustic Inversion Using Prior Information on Range-Dependent Seabed Layering," *IEEE J. Oceanic Eng.*, vol. 47, pp. 594–606, July 2022.

[9] C. Yardim, P. Gerstoft, and W. S. Hodgkiss, "Geoacoustic and source tracking using particle filtering: Experimental results," *J. Acoust. Soc. Am.*, vol. 128, pp. 75–87, July 2010.

[10] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *J. Acoust. Soc. Am.*, vol. 152, pp. 107–151, July 2022.

[11] A. Weiss, A. C. Singer, and G. W. Wornell, "Towards Robust Data-Driven Underwater Acoustic Localization: A Deep CNN Solution with Performance Guarantees for Model Mismatch," in *Proc. IEEE ICASSP*, pp. 1–5, June 2023.

[12] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, "Taking the Human Out of the Loop: A Review of Bayesian Optimization," *Proc. IEEE*, vol. 104, pp. 148–175, Jan. 2016.

[13] S. Greenhill, S. Rana, S. Gupta, P. Vellanki, and S. Venkatesh, "Bayesian Optimization for Adaptive Experimental Design: A Review," *IEEE Access*, vol. 8, pp. 13937–13948, 2020.

[14] M. Kaselimi, N. Doulamis, A. Doulamis, A. Voulodimos, and E. Protopapadakis, "Bayesian-optimized Bidirectional LSTM Regression Model for Non-intrusive Load Monitoring," in *Proc. IEEE ICASSP*, pp. 2747–2751, May 2019.

[15] H. B. Moss, V. Aggarwal, N. Prateek, J. Gonzalez, and R. Barra-Chicote, "BOFFIN TTS: Few-Shot Speaker Adaptation by Bayesian Optimization," in *Proc. IEEE ICASSP*, pp. 7639–7643, May 2020.

[16] R. Goel, F. Abhimanyu, K. Patel, J. Galeotti, and H. Choset, "Autonomous Ultrasound Scanning using Bayesian Optimization and Hybrid Force Control," in *Proc. IEEE Int. Conf. Robot. Autom.*, pp. 8396–8402, May 2022.

[17] K. D. Polyzos, Q. Lu, and G. B. Giannakis, "Bayesian Optimization with Ensemble Learning Models and Adaptive Expected Improvement," in *Proc. IEEE ICASSP*, pp. 1–5, June 2023.

[18] W. F. Jenkins II, P. Gerstoft, and Y. Park, "Bayesian optimization with Gaussian process surrogate model for source localization," *J Acoust. Soc. Am.*, vol. 154, pp. 1459–1470, Sept. 2023.

[19] F. B. Jensen, W. A. Kuperman, M. B. Porter, and H. Schmidt, *Computational Ocean Acoustics*. Modern Acoustics and Signal Processing, New York, NY: Springer New York, 2011.

[20] M. B. Porter, "The KRAKEN normal mode program," *SACLANT Undersea Research Centre Memorandum (SM-245)/Naval Research Laboratory Memorandum Report 6920*, Sept. 1991.

[21] K. P. Murphy, *Probabilistic Machine Learning: An Introduction*. Cambridge, MA: MIT Press, 2022.

[22] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A Limited Memory Algorithm for Bound Constrained Optimization," *SIAM J. Sci. Comput.*, vol. 16, no. 5, pp. 1190–1208, 1995.

[23] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimization," *ACM Trans. Math. Software*, vol. 23, pp. 550–560, Dec. 1997.

[24] D. R. Jones, "A Taxonomy of Global Optimization Methods Based on Response Surfaces," *J. Global Optim.*, vol. 21, no. 4, pp. 345–383, 2001.

[25] M. Balandat, B. Karrer, D. R. Jiang, S. Daulton, B. Letham, A. G. Wilson, and E. Bakshy, "BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 21524–21538, 2020.

[26] I. M. Sobol', "On the distribution of points in a cube and the approximate evaluation of integrals," *USSR Comp. Math. and Math. Phys.*, vol. 7, pp. 86–112, Jan. 1967.

[27] Marine Physical Laboratory, "SWellEx-96 Experiment." http://swellex96.ucsd.edu/.

[28] R. T. Bachman, P. W. Schey, N. O. Booth, and F. J. Ryan, "Geoacoustic databases for matched-field processing: Preliminary results in shallow water off San Diego, California," *J. Acoust. Soc. Am.*, vol. 99, pp. 2077–2085, Apr. 1996.

[29] G. L. D'Spain, J. J. Murray, and W. S. Hodgkiss, "Mirages in shallow water matched field processing," *J. Acoust. Soc. Am.*, vol. 105, pp. 3245–3265, June 1999.

Chapter 6

Geoacoustic inversion with Bayesian optimization

Geoacoustic inversion is a computationally expensive task in high-dimensional parameter spaces, typically requiring thousands of evaluations of a forward model to estimate the geoacoustic environment, In this study, we demonstrate Bayesian optimization (BO), an efficient global optimization method that is capable of estimating geoacoustic parameters in 7-dimensional space within hundreds of evaluations instead of thousands. BO iteratively searches parameter space for the global optimum of an objective function, defined in this study as the Bartlett power. Each step consists of fitting a Gaussian process (GP) surrogate model to observed data, and then choosing a new point to evaluate using a heuristic acquisition function. The ideal acquisition function balances exploration of the parameter space in regions with high uncertainty with exploitation of high-performing regions. In this study, two acquisition functions are evaluated: upper confidence bound and expected improvement. BO is demonstrated for both simulated and experimental data from a shallow-water environment. Results indicate BO rapidly estimates optimal parameters compared to quasi-random search.

6.1 Introduction

Geoacoustic inversion is typically a computationally expensive task, requiring thousands of evaluations of a forward model to estimate underlying parameter distributions. While compu-

tational costs have been mitigated through advances in computing capabilities, settings remain in which efficient estimates of the ocean environment are desirable. For example, the proliferation of battery-limited autonomous underwater vehicles has led to a demand in *in-situ* estimates of the ocean environment to improve acoustical navigation, as sound propagation is inextricably linked to the environment. This study proposes Bayesian optimization (BO) as a sample-efficient method for performing geoacoustic inversion.

In the forward problem, underwater acoustic propagation models provide reliable predictions of sound fields in underwater environments. In addition to source-receiver geometry, an important factor in the accuracy of these models is the inclusion of sound speed profile (SSP) and seabed properties, which affect the acoustic field in water due to refraction and interactions with the bottom [1]. While SSP in the water column can be measured using conductivity-temperature-depth (CTD) instruments, geoacoustic parameters are far more challenging and expensive to measure, particularly when the seabed consists of inhomogeneous geologic layers. Geoacoustic inversion is therefore an important field of underwater acoustics, providing feasible constraints on environmental models for parameters that are difficult to directly sample.

The inverse problem, i.e., estimating model parameters given a propagation model and observed data, is conventionally solved through matched field processing (MFP), which evaluates samples from the parameter space with the objective of finding parameters that yield a predicted acoustic field that matches the observed data [2, 3]. These matches, or correlations, comprise an ambiguity surface that is computed over the parameter space. Numerous implementations of geoacoustic inversion have been developed in the past decades. For two- or three-dimensional parameter space, such as with source localization, MFP is typically implemented as grid search [2], where parameters like source range and depth are discretized into equally spaced resolution cells according to physical characteristics of the waveguide [4]. For geoacoustic inversion, grid search—even with coarse resolution—is computationally unfeasible due to the high dimensionality of the parameter space, as computational cost scales exponentially with dimensionality.

Geoacoustic parameter estimation in high dimensions became computationally feasible

and statistically tractable with the adoption of Markov chain Monte Carlo (MCMC) sampling algorithms. By treating the ambiguity surface as a posterior distribution over the parameter space, simulated annealing [3], genetic algorithms [5], and Gibbs sampling [6] provide statistical estimates of the parameter distributions. Hybrid methods combine these distributions with gradient-based optimization methods to fine-tune the estimates [7, 8]. For time-evolving cases, particle filtering was demonstrated for geoacoustic inversion during an acoustic source tow [9]. In cases where the number of geoacoustic parameters is unknown (for example, the number and makeup of sediment layers in the seabed), transdimensional geoacoustic inversion provides estimates of the number of geoacoustic parameters, in addition to their distributions [10]. For a detailed history and review of geoacoustic inversion methods, please consult [11]. Recent developments in geoacoustic inversion involve joint estimation of geoacoustic parameters and SSP in the water column using transdimensional inversion [12]. Additionally, data-driven methods using neural networks have been demonstrated for SSP and geoacoustic inversion [13, 14, 15, 16, 17]. Further discussion of advances in machine learning methods related to geoacoustic inversion can be found in [18].

One of the drawbacks of MCMC is its computational cost: thousands of evaluations of the forward model are required to form parameter estimates, and methods like genetic algorithms have tunable parameters that require optimization [7, 19, 20, 21]. Machine learning-based approaches are also computationally expensive, requiring ample data for training and prediction and numerous runs for hyperparameter optimization Training can take many hours, and models are not readily transferred to other geoacoustic environments, necessitating retraining on new data when models can not be generalized.

To overcome these challenges, we demonstrate Bayesian optimization (BO), a sample-efficient, global optimization strategy which adaptively chooses which samples to evaluate according to previous evaluations of the ambiguity surface [22, 23]. Treating the ambiguity surface as an objective function, BO seeks to find the global minimum. At every step of the optimization, BO fits a Gaussian process [24] (GP) surrogate model to the ambiguity surface; the

uncertainty encoded in the GP is then used by a heuristic acquisition function to suggest which point in parameter space to evaluate next [25]. BO requires no information about the gradient of the ambiguity surface, and though not invulnerable to converging on local optima, its adaptive nature makes it more robust for multimodal objective function optimization than gradient-based methods. Bayesian optimization was previously demonstrated for source localization in [26], where a more detailed comparison to other optimization strategies can be found.

The paper is organized as follows: Sec. 6.2 presents the geoacoustic inversion problem; Sec. 6.3 describes the Bayesian optimization framework; Sec. 6.4 summarizes the data and geoacoustic environment model used for simulated and experimental data analysis; Sec. 6.5 presents analysis of BO results using simulated and experimental data; and Sec. 6.6 provides a discussion of the BO results.

6.2 Inversion framework

6.2.1 Parameterization

Consider an acoustic pressure field $\mathbf{q}(\omega_l) \in \mathbb{C}^J$ sampled by a vertical line array (VLA) with J elements processed at frequencies $\mathbf{\Omega} = [\omega_1, \ldots, \omega_L]$. Observed data are the sum of predictions of the data $\mathbf{p}(\mathbf{m}, \omega_l) \in \mathbb{C}^J$ given a model $\mathbf{m} \in \mathscr{M}^D$ with D parameters, and an error term $\mathbf{e}(\omega_l)$:

$$\mathbf{q}(\omega_l) = \mathbf{p}(\mathbf{m}, \omega_l) + \mathbf{e}(\omega_l). \tag{6.1}$$

Predicted data are modeled by:

$$\mathbf{p}(\mathbf{m}, \omega_l) = \mathbf{w}(\mathbf{m}, \omega_l) S(\omega_l) \tag{6.2}$$

where $\mathbf{w}(\mathbf{m}, \omega_l) \in \mathbb{C}^J$ is a replica pressure field produced by an acoustic propagation model, and $S(\omega_l) \in \mathbb{C}$ is an unknown deterministic source term. For brevity, we denote frequency dependence by $\mathbf{q}_l = \mathbf{q}(\omega_l)$, etc. Replica pressure fields $\mathbf{w}_l(\mathbf{m})$ are generated using the KRAKEN

normal mode propagation model [27].

6.2.2 Objective function

The estimated model $\widehat{\mathbf{m}}$ is obtained by matching $\mathbf{q}_l$ and $\mathbf{p}_l$ through minimization of an objective function ϕ:

$$\widehat{\mathbf{m}} = \arg\min_{\mathscr{M}} [\phi(\mathbf{m})] \tag{6.3}$$

$$\widehat{\phi} = \phi(\widehat{\mathbf{m}}). \tag{6.4}$$

ϕ is constructed from the Bartlett objective function [28, 29]:

$$\phi(\mathbf{m}) = \prod_{l=1}^{L} \left(\text{tr}\, \widehat{\mathbf{R}}_l - \frac{\mathbf{w}_l^{\mathsf{H}}(\mathbf{m}) \widehat{\mathbf{R}}_l \mathbf{w}_l(\mathbf{m})}{\mathbf{w}_l^{\mathsf{H}}(\mathbf{m}) \mathbf{w}_l(\mathbf{m})} \right) \tag{6.5}$$

where $\widehat{\mathbf{R}}_l \in \mathbb{C}^{J \times J}$ is the sample covariance matrix (SCM) from K snapshots:

$$\widehat{\mathbf{R}}_l = \frac{1}{K} \sum_{k=1}^{K} \mathbf{q}_{k,l} \mathbf{q}_{k,l}^{\mathsf{H}}. \tag{6.6}$$

6.3 Bayesian optimization

In BO, a GP surrogate model $\mathscr{GP}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ approximates the objective function $\phi(\mathbf{m})$ by performing GP regression over a set of observed data $\mathscr{D}$. Next, an acquisition function takes the GP surrogate model as its input and suggests a new point in $\mathscr{M}$ to evaluate. The process repeats as data are added by exploration of $\mathscr{M}$ until a fixed budget of N_{total} evaluations, or trials, is exhausted.

6.3.1 Gaussian process surrogate model

Here we follow the derivations of [30]. A Gaussian process is a collection of N random points:

$$\mathbf{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_N] \in \mathscr{M}^{D \times N}, \tag{6.7}$$

any finite number of which have a joint Gaussian distribution. Given a real process $\mathbf{f} = [\phi(\mathbf{m}_1), \ldots, \phi(\mathbf{m}_N)]^\mathsf{T}$, a GP is described by a mean function,

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{f}] = [\mu(\mathbf{m}_1), \ldots, \mu(\mathbf{m}_N)]^\mathsf{T} \in \mathbb{R}^N, \tag{6.8}$$

where $\mu(\mathbf{m}_n)$ is the mean at $\mathbf{m}_n$; and a covariance function,

$$\boldsymbol{\Sigma}_{ij} = \mathbb{E}[(f(\mathbf{m}_i) - \mu(\mathbf{m}_i))(f(\mathbf{m}_j) - \mu(\mathbf{m}_j))] \tag{6.9}$$

$$= \mathscr{K}(\mathbf{m}_i, \mathbf{m}_j) \in \mathbb{R}^{N \times N}, \tag{6.10}$$

where $\mathscr{K}(\mathbf{m}_i, \mathbf{m}_j)$ is a kernel function measuring the similarity between points $\mathbf{m}_i$ and $\mathbf{m}_j$. The covariance function is assumed diagonal such that the variance at a point $\sigma(\mathbf{m})$ is obtained from $\boldsymbol{\Sigma} \approx \text{diag}\left([\sigma^2(\mathbf{m}_1), \ldots, \sigma^2(\mathbf{m}_N)]\right)$. The real process $\mathbf{f}$ is modeled by the GP as $\mathbf{f} \sim \mathscr{GP}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and has observations comprising parameters evaluated by the objective function:

$$\mathscr{D} = \{(\mathbf{m}_n, y_n),\ n = 1 : N\} = \{\mathbf{M}, \mathbf{y}\}, \quad \mathbf{y} \in \mathbb{R}^N. \tag{6.11}$$

The GP predicts N^* unobserved outputs $\mathbf{f}_*$ at inputs $\mathbf{M}_{*,D \times N^*} = [\mathbf{m}_1^*, \ldots, \mathbf{m}_{N^*}^*]$. From [30, eq. 17.33], the joint distribution of the observed process $\mathbf{y}$ and the predictive distribution $\mathbf{f}_*$ is:

$$p(\mathbf{y}, \mathbf{f}_* | \mathbf{M}, \mathbf{M}_*) = \begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} = \mathscr{N}\left(\begin{bmatrix} \boldsymbol{\mu}_M \\ \boldsymbol{\mu}_* \end{bmatrix}, \begin{bmatrix} \widehat{\mathbf{K}}_{M,M} & \mathbf{K}_{M,*} \\ \mathbf{K}_{M,*}^\mathsf{T} & \mathbf{K}_{*,*} \end{bmatrix} \right) \tag{6.12}$$

where $\boldsymbol{\mu}_M$ and $\boldsymbol{\mu}_*$ are the mean functions at $\mathbf{M}$ and $\mathbf{M}_*$, respectively; and

$$\widehat{\mathbf{K}}_{M,M} = \mathbf{K}_{M,M} + \sigma_y^2 \mathbf{I} = \mathscr{K}(\mathbf{M},\mathbf{M})^{N\times N} + \sigma_y^2 \mathbf{I} \tag{6.13}$$

$$\mathbf{K}_{M,*} = \mathscr{K}(\mathbf{M},\mathbf{M}_*)^{N\times N_*} \tag{6.14}$$

$$\mathbf{K}_{*,*} = \mathscr{K}(\mathbf{M}_*,\mathbf{M}_*)^{N_*\times N_*}. \tag{6.15}$$

From [30, eq. 17.34], the GP is conditioned on the new observations, giving the posterior distribution:

$$p(\mathbf{f}_*|\mathscr{D},\mathbf{M}_*) = \mathscr{N}(\mathbf{f}_*|\boldsymbol{\mu}_{*|M},\boldsymbol{\Sigma}_{*|M}) \tag{6.16}$$

$$\boldsymbol{\mu}_{*|M} = \boldsymbol{\mu}_* + \mathbf{K}_{M,*}^{\mathsf{T}} \widehat{\mathbf{K}}_{M,M}^{-1}(\mathbf{y} - \boldsymbol{\mu}_M) \tag{6.17}$$

$$\boldsymbol{\Sigma}_{*|M} = \mathbf{K}_{*,*} - \mathbf{K}_{M,*}^{T} \widehat{\mathbf{K}}_{M,M}^{-1} \mathbf{K}_{M,*}. \tag{6.18}$$

The kernel function (6.10) measures similarity between two points in $\mathscr{M}$. We use a positive-definite and stationary kernel with real-valued inputs, i.e.,

$$\mathscr{K}(\mathbf{m}_i,\mathbf{m}_j) = \mathscr{K}(\mathbf{r}), \quad \mathbf{r} = \mathbf{m}_i - \mathbf{m}_j \tag{6.19}$$

and specifically adopt the Matern kernel with smoothness parameter $\nu = 5/2$:

$$\mathscr{K}\left(\mathbf{r};\frac{5}{2},\mathbf{l}\right) = \sigma_y^2 \prod_{d=1}^{D} \left(1 + \frac{\sqrt{5}r_d}{l_d} + \frac{5r_d^2}{3l_d^2}\right) \exp\left(-\frac{\sqrt{5}r_d}{l_d}\right) \tag{6.20}$$

where $\sigma_y \in \mathbb{R}$ and $l_d \in \mathbb{R}$ are hyperparameters, with σ_y^2 estimating the noise variance of the GP and $\mathbf{l} = [l_1,\ldots,l_D]$ controlling the length scale in dimension d [30, eq. 17.13].

Kernel hyperparameters $\boldsymbol{\theta} = [\sigma_y^2,\mathbf{l}]$ must be optimized for the GP surrogate model to appropriately reflect the data and is efficiently performed with an empirical Bayes approach.

Noting $\boldsymbol{\theta}$ is implicit in $\widehat{\mathbf{K}}_{M,M}$, marginalizing the product of the Gaussians

$$p(\mathbf{f}|\mathbf{M},\boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}_M, \widehat{\mathbf{K}}_{M,M}) \tag{6.21}$$

$$p(\mathbf{y}|\mathbf{f},\mathbf{M}) = \prod_{n=1}^{N} \mathcal{N}(y_n|f_n, \sigma_y^2). \tag{6.22}$$

with respect to $\mathbf{f}$ results in the marginal likelihood, itself a Gaussian [30, eq. 17.51]:

$$p(\mathbf{y}|\mathbf{M},\boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f},\mathbf{M},\boldsymbol{\theta}) p(\mathbf{f}|\mathbf{M},\boldsymbol{\theta}) d\mathbf{f} \tag{6.23}$$

$$= \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_M, \widehat{\mathbf{K}}_{M,M}). \tag{6.24}$$

To find the optimal $\boldsymbol{\theta}$, the log marginal likelihood is maximized [30, eq. 17.51]:

$$\begin{aligned}\mathscr{L} &= \log p(\mathbf{y}|\mathbf{M},\boldsymbol{\theta}) = \log \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_M, \widehat{\mathbf{K}}_{M,M}) \\ &= -\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_M)^{\mathsf{T}} \widehat{\mathbf{K}}_{M,M}^{-1}(\mathbf{y}-\boldsymbol{\mu}_M) \\ &\quad -\frac{1}{2}\log|\widehat{\mathbf{K}}_{M,M}| - \frac{N}{2}\log(2\pi)\end{aligned} \tag{6.25}$$

$$\frac{\partial \mathscr{L}}{\partial \theta_j} = \frac{1}{2}\mathrm{tr}\left[(\boldsymbol{\gamma}\boldsymbol{\gamma}^{\mathsf{T}} - \widehat{\mathbf{K}}_{M,M})\frac{\partial \widehat{\mathbf{K}}_{M,M}}{\partial \theta_j}\right], \tag{6.26}$$

where $\boldsymbol{\gamma} = \widehat{\mathbf{K}}_{M,M}^{-1}(\mathbf{y}-\boldsymbol{\mu}_M)$ [30, eq. 17.52]. Hyperparameter optimization is performed using 50 steps of AdamW, an adaptive gradient-based stochastic optimizer [31].

The empirical Bayes hyperparameter optimization of (6.20) is a form of automatic relevancy determination (ARD) [30, sec. 17.1.2.1]. During optimization of an ARD kernel, if a particular parameter does not affect the value of the objective function, the associated hyperparameter for that parameter's dimension will approach $l_d = \infty$.

6.3.2 Acquisition function

Bayesian optimization is performed sequentially over a fixed budget of N_{total} trials. At trial t, a point $\mathbf{m}_t$ is evaluated by the objective function $\phi(\mathbf{m}_t)$ (6.5). The next point $\mathbf{m}_{t+1}$ is suggested through optimization of a heuristic acquisition function $\alpha(\phi(\mathbf{m}))$:

$$\mathbf{m}_{t+1} = \underset{\mathbf{m}\in\mathscr{M}}{\arg\max}\left[\alpha\left(\phi(\mathbf{m})\right)\right] \tag{6.27}$$

The ideal acquisition function balances the exploration of regions of high uncertainty with the exploitation of well-performing regions. Though there are many acquisition functions in the literature, we examine the upper confidence bound (UCB) and expected improvement (EI) functions for their robust performance and simplicity.

Upper confidence bound

The UCB acquisition function [25, 32] is the weighted sum of the GP posterior mean function $\mu(\mathbf{m})$ and uncertainty $\sigma(\mathbf{m})$:

$$\alpha_{\text{UCB}}(\phi(\mathbf{m})) = \mu(\mathbf{m}) + \kappa\sigma(\mathbf{m}). \tag{6.28}$$

κ is a tunable parameter that controls the tradeoff between exploration and exploitation. UCB is sensitive to the choice of κ and requires tuning to achieve optimal performance [25]. Initially, UCB is dominated by the uncertainty term $\sigma(\mathbf{m})$ and favors exploration of the parameter space. As more data are added, the GP mean function $\mu(\mathbf{m})$ dominates and UCB exploits the best-performing regions.

Expected improvement

The EI acquisition function requires no hyperparameter tuning, adaptively balancing exploitation with exploration [25, 23]. EI quantifies the expected amount of improvement a point is expected to yield over the best previously observed objective function $\widehat{\phi}$ (6.4) and is defined

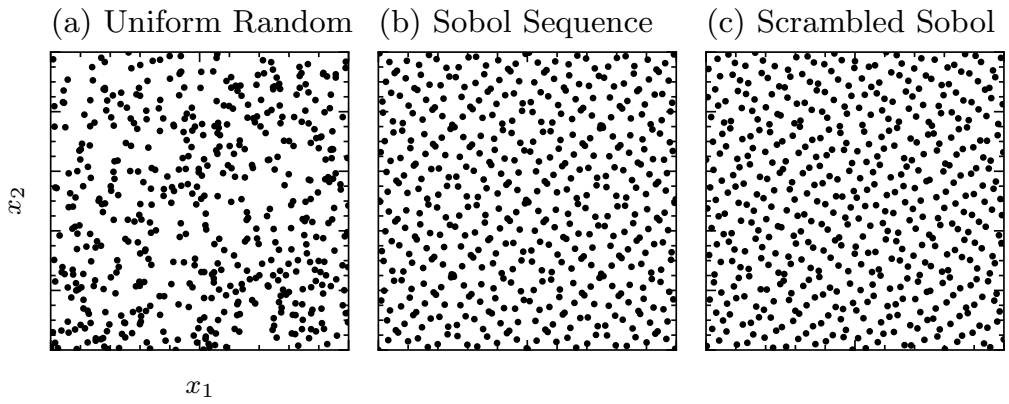

Figure 6.1. 512 points drawn from a (a) uniform distribution, (b) Sobol sequence, and (c) scrambled Sobol sequence.

as:

$$\begin{aligned} \alpha_{\text{EI}}(\phi(\mathbf{m})) &= \mathbb{E}\left[\max(\phi(\mathbf{m}) - \widehat{\phi}, 0)\right] \\ &= \sigma(\mathbf{m})(z\Phi(z) + \varphi(z)) \end{aligned} \tag{6.29}$$

where z is the standardized improvement:

$$z = \frac{\widehat{\phi} - \mu(\mathbf{m})}{\sigma(\mathbf{m})}, \tag{6.30}$$

Φ is the cumulative distribution function, and φ is the normal probability distribution function.

6.3.3 Implementation

To establish a reasonable prior over the objective function, the GP surrogate model is initialized with N_{init} warmup trials. For this study, warm-up trials are generated using quasi-random sampling. A Sobol sequence is a quasi-random sequence that generates points in $\mathscr{M}$ such that the points are spread more evenly than when drawn from a uniform distribution [33]. This study uses a variation of a Sobol sequence: the scrambled Sobol sequence randomizes the order of the points, which can reduce the correlation between dimensions [34]. Figure 6.1 illustrates the difference between uniform, Sobol sequence, and scrambled Sobol sequence sampling.

Table 6.1. Pseudocode for Bayesian optimization.

Input: Parameter domain $\mathscr{M}$, objective function ϕ, kernel function $\mathscr{K}$, acquisition function α, total trials N_{total}.
Output: Best estimate of parameters $\widehat{\mathbf{m}}$

$\widehat{\mathbf{m}}, \widehat{\phi} \leftarrow \text{SOBOL}[\mathscr{M}, N_{\text{init}}]$
for $t = N_{\text{init}} + 1$ **to** N_{total} **do**
  $\boldsymbol{\mu} \leftarrow \mathbb{E}[\phi(\mathbf{M})]$ [Eq. (6.8)]
  $\boldsymbol{\theta} \leftarrow \text{ADAMW}\,[\mathscr{K}(\mathbf{M}, \mathbf{M}; \boldsymbol{\theta})]$ [Eqs. (6.25)-(6.26)]
  $\boldsymbol{\Sigma} \leftarrow \mathscr{K}(\mathbf{M}, \mathbf{M}; \boldsymbol{\theta})$ [Eq. (6.10)]
  $\mathbf{m}_t \leftarrow \arg\max_{\mathbf{m}} \alpha(\phi(\mathbf{m}))$ [Sec. 6.3.2]
  $\phi_t \leftarrow \phi(\mathbf{m}_t)$
  if $\phi_t > \widehat{\phi}$ **then**
    $\widehat{\mathbf{m}} \leftarrow \mathbf{m}_t$, $\widehat{\phi} \leftarrow \phi_t$

BO is implemented using the BOTORCH Python library [35], with pseudocode given in Table 6.1 and algorithm parameters in Table 6.2. The GP surrogate model is initialized with N_{init} warmup trials generated by a quasi-random Sobol sequence [33] of points from the parameter space $\mathscr{M}$. GP regression is performed on the observed data $\mathscr{D}$ and the next trial is suggested by (6.27); this repeats until a fixed budget of N_{total} trials is expended. Since the acquisition function can be non-convex but is inexpensive to evaluate, (6.27) is optimized through quasi-random Monte Carlo sampling: for N_{restart} random restarts, the acquisition function is evaluated at N_{acq} points drawn from a Sobol sequence and maximized using L-BFGS-B [36, 37]. Values for N_{restart} and N_{init} are given in Table 6.2.

To improve numerical stability while fitting the GP surrogate model, data are transformed by normalizing model parameters $\mathbf{m}$ to $[0,1]$ and standardizing objective function values $\mathbf{y}$ to zero mean and unit variance. Hyperparameter optimization is improved by assigning prior distributions to the kernel hyperparameters: kernel length scale $\mathbf{l}$ values are drawn from a Gamma distribution with shape $a = 3$ and rate $b = 6$, giving a mean of 0.5; and noise variance σ_y^2 values are drawn from a Gamma distribution with shape $a = 2$ and rate $b = 0.15$.

One iteration of BO for a one-dimensional ambiguity surface over source range is shown in Fig. 6.2 with the GP surrogate model for the objective function $\phi(\mathbf{m})$ and the corresponding

Table 6.2. Bayesian optimization algorithm parameters.

Parameter	Description	Value
N_{total}	Total trials	500
N_{init}	Warm-up trials	200
N_{acq}	Samples for acquisition function optimization	1024
N_{restart}	Acquisition function re-starts	40

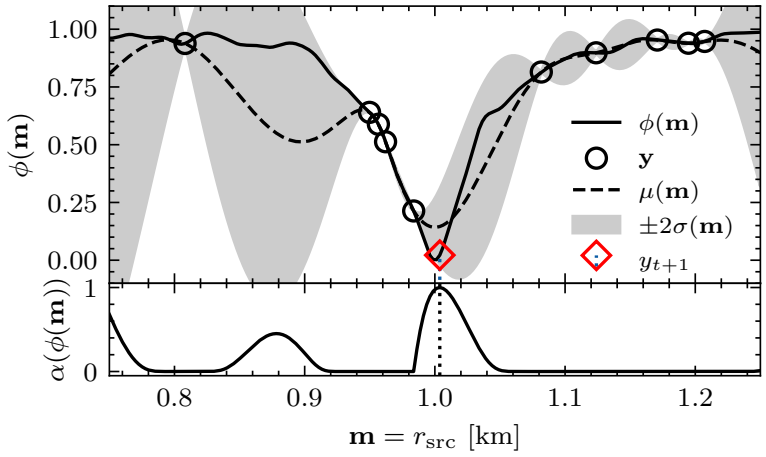

Figure 6.2. Gaussian process regression (upper panel) of the objective function $\phi(\mathbf{m})$ for one-dimensional ambiguity surface over source range r_{src}. The true surface (solid) is approximated by the mean function $\mu(\mathbf{m})$ (dashed) and uncertainty $\sigma(\mathbf{m})$ (shaded) conditioned on observed data $\mathbf{y}$ (dots). The next point y_{t+1} is obtained from the maximum of the acquisition function $\alpha(\mathbf{m})$ (lower panel), shown here normalized to $[0,1]$.

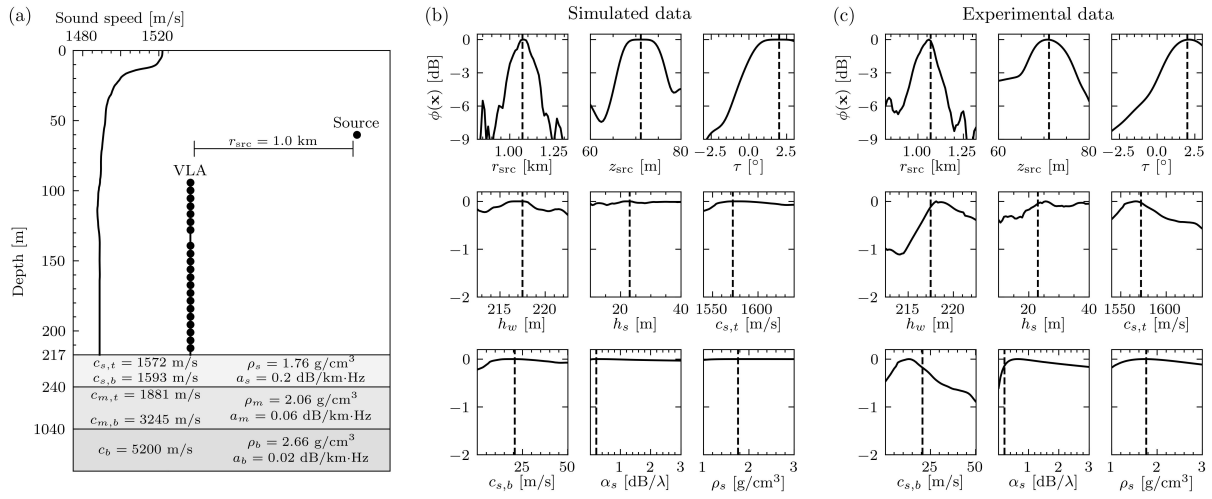

Figure 6.3. (a) Environmental model used for simulations and geoacoustic inversion. Sensitivity analyses for (b) simulated and (c) experimental data; parameter estimates are indicated with vertical dashed lines.

acquisition function $\alpha(\phi(\mathbf{m}))$. In this example, the GP surrogate model was fit to 10 samples from the objective function. The EI acquisition function is non-zero in regions of high uncertainty and low objective function value, and zero in regions of low uncertainty and high objective function value. The next sample is suggested by the maximum of the acquisition function.

6.4 Data and environment

BO is demonstrated using data from the SWellEx-96 experiment, conducted off the coast of southern California in shallow water [38]. Data were recorded on a 64-element vertical line array (VLA) deployed in 217 m of water with elements spaced evenly between 94.125 m and 212.25 m. The sampling rate of the VLA was 1.5 kHz. During event S5, R/V *Sproul* towed an acoustic source from south to north at a speed of 5 knots and depth of 60 m, with the closest point of approach (CPA) to the VLA occurring 1 km to the east of the array. The source transmitted a comb signal comprising 13 tones between 49 and 388 Hz. Hann-windowed time series data are processed from 21 of the 64 channels in 8,192-sample (2.7 s) segments with NFFT $= 8,192$ at $\mathbf{\Omega} = [148, 235, 388]$ Hz. $K = 8$ segments with 50% overlap are used to form the SCM (6.6).

Though water depths at the source range and VLA differ by as much as 40 m, a range-

Table 6.3. Model **m** parameterization.

Parameter	Definition	Bounds
r_{src} [km]	Source range	[0.75, 1.25]
z_{src} [m]	Source depth	[60, 80]
τ [$^\circ$]	Array tilt	[-3, 3]
h_w [m]	Water depth	[212, 222]
h_s [m]	Sediment thickness	[10, 40]
$c_{s,t}$ [m/s]	Sediment top sound speed	[1500, 1800]
$c_{s,b}$ [m/s]	Sediment bottom sound speed	[1.5, 2.5]

independent model is adopted to predict the acoustic field at the array. Range dependence due to differing bottom depths at the source and VLA results in the source appearing farther and deeper than it is, and is accounted for through straightforward corrections [39]. At CPA, any discrepancy between *Sproul*'s GPS range and the source's true range as a result of tow-cable scope is negligible, since the array is nearly broadside to the ship-source axis at CPA.

The SWellEx-96 geoacoustic and oceanographic environments are well characterized [40, 38] and serve as a useful testbed for BO. The environment model is depicted in Fig. 6.3a, where the geoacoustic environment is parameterized by two layers of sediment (subscript s) and mudrock (subscript m) situated atop a bedrock halfspace (subscript b), and the water column consists of a downward refracting sound speed profile (SSP). Model parameters **m** and boundaries are listed in Table 6.3.

Underwater acoustic propagation depends on many parameters, not all of which contribute equally to the acoustic field. To determine approximately which parameters are most important, we perform a sensitivity analysis by sweeping through one parameter in **m** at a time while fixing all others at their anticipated values. Sensitivity analyses for simulated and experimental data are shown in Fig. 6.3b and Fig. 6.3c, respectively. Though a one-dimensional sensitivity analysis may fail to fully convey higher-dimensional structure in the data, it is a useful tool for coarsely identifying which parameters affect the acoustic field most strongly. From Figs. 6.3b and 6.3c, the acoustic field is most sensitive to source range r_{src}, source depth z_{src}, and array tilt τ. Simulations indicate the remaining parameters show little sensitivity, but experimental data

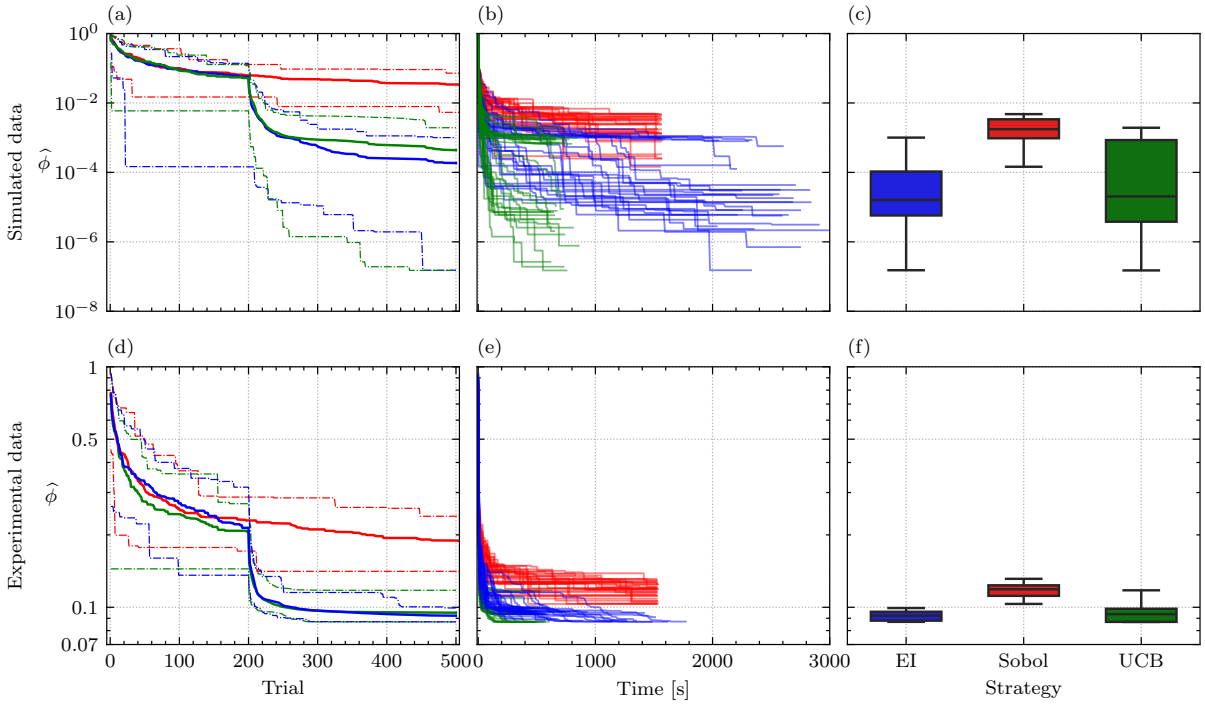

Figure 6.4. Lowest observed values of $\widehat{\phi}$ from 30 Monte Carlo runs for (a)-(c) simulated and (d)-(f) experimental data. (a)(d) $\widehat{\phi}$ vs. trial; solid lines indicate the mean value of the Monte Carlo runs at that trial, and dashed lines indicate minimum and maximum values. (b)(e) $\widehat{\phi}$ vs. wall time; each trace represents a Monte Carlo run. (c)(f) Distribution of final values of $\widehat{\phi}$; outer horizontal lines represent minimum and maximum values and inner horizontal lines denote quartile boundaries.

reveal sensitivity to water depth h_w, sediment thickness h_s, and sound speeds $c_{s,t}$, $c_{s,b}$ at the top and bottom of the sediment layer, respectively. Attenuation and density show little sensitivity at the processed frequencies in the given environment and are omitted from the inversion.

6.5 Example

Bayesian optimization is demonstrated on both simulated and experimental data using the UCB and EI acquisition functions. Sobol sequence sampling is also performed to serve as a comparison. Since BO takes more time per trial than quasi-random sampling, we use a Sobol sequence with 50,000 points. Since each strategy is sensitive to initialization, 30 Monte Carlo runs are performed to characterize performance.

Figures 6.4a and 6.4d show the lowest observed values of $\widehat{\phi}$ (6.4) for simulated and experimental data, respectively. Solid lines indicate the mean value of $\widehat{\phi}$ vs. trial, and dashed lines indicate the minimum (best) and maximum (worst) value vs. trial. The BO strategies are run with $N_{\text{total}} = 500$ trials and initialized with $N_{\text{init}} = 200$ warmup trials. With the GP surrogate model conditioned on the warmup trials, BO rapidly locates optimal regions to evaluate. Both the UCB and EI acquisition functions yield superior optimization performance to Sobol sampling given the same trial budget, with EI providing slightly better results than UCB.

In practice, a grid search or quasi-random search would not be restricted to $N_{\text{total}} = 500$ trials. While the per-trial comparison of Figs. 6.4a and 6.4d are useful, equally important to consider is the wall time that elapses to run each strategy. To examine whether the time-per-trial of BO outweighs running grid or quasi-random search with a large number of points, Figs. 6.4b and 6.4e show $\widehat{\phi}$ vs. elapsed wall time for each Monte Carlo run of the BO and Sobol strategies. BO with the UCB and EI acquisition functions is shown with $N_{\text{init}} = 200$ and $N_{\text{total}} = 500$, while Sobol sampling is evaluated for $N_{\text{total}} = 50,000$ trials, which takes a similar amount of time to run as 500 trials of BO. BO achieves better values for $\widehat{\phi}$ than Sobol sampling even when the latter is permitted to evaluate two orders of magnitude more trials. Since UCB is a computationally simpler acquisition function to optimize, its performance vs. wall time is particularly noteworthy, rapidly converging on low values of $\widehat{\phi}$. The more complicated optimization of the EI acquisition function takes more time but yields superior results to UCB.

The distributions of final optimization results from each strategy's 30 Monte Carlo runs are shown in Figs. 6.4c and 6.4f. Distributions consist of values for $\widehat{\phi}$ at the 500^{th} trial for the BO strategies and the $50{,}000^{\text{th}}$ trial for Sobol search. With both simulated and experimental data, BO with the EI acquisition function achieves the best values for $\widehat{\phi}$. BO with the UCB acquisition function also achieves improved performance over Sobol sampling, but the distribution of final values is wider than for EI since UCB can get stuck in local optima and is unable to adaptively balance between exploitation and exploration like EI [25, 32].

Figure 6.5 shows histograms of parameter estimates from the 30 Monte Carlo runs of

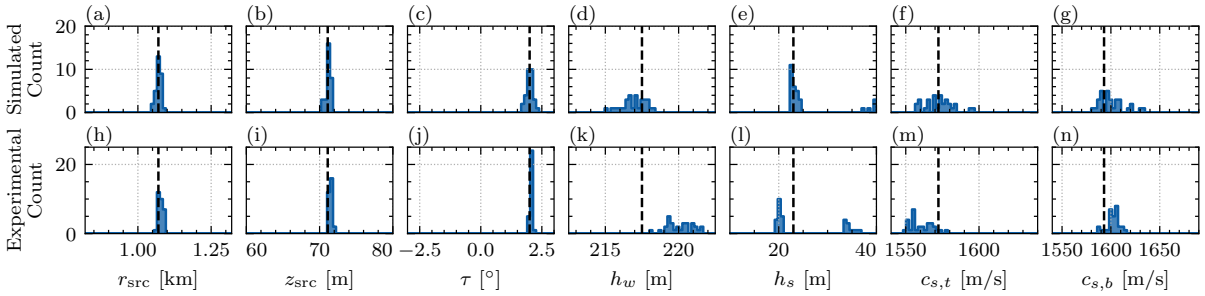

Figure 6.5. Histograms of parameter estimates from 30 Monte Carlo runs of Bayesian optimization with expected improvement acquisition function and $N_{\mathrm{init}} = 200$. Estimates are shown for (a)-(g) simulated and (h)-(n) experimental data; true and expected values are indicated by the black dashed line for simulated and experimental data, respectively.

BO with the EI acquisition function and $N_{\mathrm{init}} = 200$ trials. For simulated data, final parameter estimates are close to the true parameter values for the most sensitive parameters identified from Fig. 6.3: source range r_{src}, source depth z_{src}, and VLA tilt τ. Estimates of water depth h_w and sediment sound speeds at the top $c_{s,t}$ and bottom $c_{s,b}$ of the first sediment layer are distributed more widely than source-receiver geometry parameters but remain consistently centered around their true values. Sediment layer thickness h_s exhibits a bimodal distribution. Results from experimental data are also consistent with anticipated values, although the less sensitive parameters exhibit biases in the estimates. Distributions of parameter estimates from both simulated and experimental data are consistent with the sensitivities shown in Fig. 6.5, indicating that the optimization can estimate the most sensitive parameters but is vulnerable to terminating in local optima in the least sensitive parameters.

6.6 Discussion

The previous analyses use $N_{\mathrm{init}} = 200$ warmup trials to initialize the GP surrogate model. This is an arbitrary selection that can be tuned according to the requirements of the optimization problem. For example, if more rapid optimization is desired, N_{init} can be set to higher values, with the remainder of the budget consisting of BO to rapidly fine-tune the solution. However, setting N_{init} too high precludes BO from exploring the parameter space and can lead to incomplete

optimization and suboptimal parameter estimates.

Because BO operates on the ambiguity surface, its performance relative to the true model parameters is dependent on the quality of the ambiguity surface. For example, in cases where there is model mismatch or noisy data, sidelobes in the ambiguity surface will become more prominent, making the surface more multimodal and increasing the likelihood that BO might converge on a local optimum. With enough noise or mismatch, sidelobes in the ambiguity surface may obfuscate the peak associated with the true parameters, and BO will estimate the wrong parameters. This dependence is not unique to BO, as any sampling algorithm that makes use of the Bartlett objective (6.5) is susceptible. Like any optimization algorithm, BO can converge on local optima if terminated too early, though various techniques have been implemented to make BO more robust in the optimization of multi-modal objective functions, including the development of robust, quasi-Monte Carlo acquisition functions, incorporation of random restart heuristics, and use of trust regions which adaptively update the parameter search space [35, 41].

One of the most flexible and intuitive aspects of BO is the ability to incorporate prior knowledge of the geoacoustic environment by placing prior distributions over the GP surrogate model hyperparameters and parameter space boundaries. For example, length scales associated with source range and depth can be constrained according to the size of the expected resolution cell [4], and parameters with little sensitivity can be assigned kernel length scale priors that favor large numbers to ensure little curvature in that dimension. Prior knowledge and expertise about the physics of propagation and the geoacoustic environment—e.g., from geologic surveys or computational models—directly inform the shape of the GP surrogate model, which is in turn used by the acquisition function to guide the search for the global optimum. The ability to specify kernel functions and hyperparameter priors in multiple dimensions enables the construction of more sophisticated GP surrogate models that can encompass this prior knowledge and has been demonstrated for high-dimensional BO problems [42].

6.7 Conclusion

Bayesian optimization (BO) efficiently finds regions of optimal performance due to its ability to incorporate observations of the objective function into its decision-making about where to sample next. With a suitably designed kernel function, the Gaussian process (GP) surrogate model provides a flexible and tractable method to incorporate prior knowledge of the geoacoustic environment into the optimization problem. The heuristic acquisition function relies on the GP to balance exploration of regions of high uncertainty with exploitation of regions where the best observed objective function values reside. The BO framework enables optimization of non-convex objective functions and requires no information about the gradient of the objective function. Using simulated and real data from a shallow-water experiment, we have demonstrated that BO strategies provide more accurate estimates of the geoacoustic parameters than quasi-random search, and do so in less time.

6.8 Acknowledgements

This research was supported by Office of Naval Research grant N00014-20-1-2555.

Chapter 6, in full, has been submitted for publication of the material as it may appear in the Journal of the Acoustical Society of America. Jenkins, William; Gerstoft, Peter; Park, Yongsung, Acoustical Society of America, 2023. The dissertation author was the primary investigator and author of this paper.

6.9 References

[1] F. B. Jensen, W. A. Kuperman, M. B. Porter, and H. Schmidt, *Computational Ocean Acoustics*. Modern Acoustics and Signal Processing, New York, NY: Springer New York, 2011.

[2] A. B. Baggeroer, W. A. Kuperman, and P. N. Mikhalevsky, "An overview of matched field methods in ocean acoustics," *IEEE J. Ocean. Eng.*, vol. 18, pp. 401–424, Oct. 1993.

[3] M. D. Collins and W. A. Kuperman, "Focalization: Environmental focusing and source localization," *J. Acoust. Soc. Am.*, vol. 90, pp. 1410–1422, Sept. 1991.

[4] S. Kim, G. F. Edelmann, W. A. Kuperman, W. S. Hodgkiss, H. C. Song, and T. Akal, "Spatial resolution of time-reversal arrays in shallow water," *J. Acoust. Soc. Am.*, vol. 110, pp. 820–829, Aug. 2001.

[5] P. Gerstoft, "Inversion of seismoacoustic data using genetic algorithms and a posteriori probability distributions," *J. Acoust. Soc. Am.*, vol. 95, pp. 770–782, Feb. 1994.

[6] S. E. Dosso, "Quantifying uncertainty in geoacoustic inversion. I. A fast Gibbs sampler approach," *J. Acoust. Soc. Am.*, vol. 111, pp. 129–142, Jan. 2002.

[7] P. Gerstoft, "Inversion of acoustic data using a combination of genetic algorithms and the Gauss–Newton approach," *J. Acoust. Soc. Am.*, vol. 97, pp. 2181–2190, Apr. 1995.

[8] M. R. Fallat and S. E. Dosso, "Geoacoustic inversion via local, global, and hybrid algorithms," *J. Acoust. Soc. Am.*, vol. 105, pp. 3219–3230, June 1999.

[9] C. Yardim, P. Gerstoft, and W. S. Hodgkiss, "Geoacoustic and source tracking using particle filtering: Experimental results," *J. Acoust. Soc. Am.*, vol. 128, pp. 75–87, July 2010.

[10] J. Dettmer, S. E. Dosso, and C. W. Holland, "Trans-dimensional geoacoustic inversion," *J. Acoust. Soc. Am.*, vol. 128, pp. 3393–3405, Dec. 2010.

[11] N. R. Chapman and E. C. Shang, "Review of Geoacoustic Inversion in Underwater Acoustics," *J. Theor. Comp. Acout.*, vol. 29, p. 2130004, Sept. 2021.

[12] S. E. Dosso and J. Bonnel, "Joint trans-dimensional inversion for water-column sound speed and seabed geoacoustic models," *JASA Express Lett.*, vol. 3, p. 060801, June 2023.

[13] M. Bianco and P. Gerstoft, "Dictionary learning of sound speed profiles," *The Journal of the Acoustical Society of America*, vol. 141, pp. 1749–1758, Mar. 2017.

[14] J. Piccolo, G. Haramuniz, and Z.-H. Michalopoulou, "Geoacoustic inversion with generalized additive models," *J. Acoust. Soc. Am.*, vol. 145, pp. EL463–EL468, June 2019.

[15] Y. Shen, X. Pan, Z. Zheng, and P. Gerstoft, "Matched-field geoacoustic inversion based on radial basis function neural network," *The Journal of the Acoustical Society of America*, vol. 148, pp. 3279–3290, Nov. 2020.

[16] M. Liu, H. Niu, Z. Li, Y. Liu, and Q. Zhang, "Deep-learning geoacoustic inversion using multi-range vertical array data in shallow water," *J. Acoust. Soc. Am.*, vol. 151, pp. 2101–2116, Mar. 2022.

[17] M. Liu, H. Niu, Z. Li, and Y. Guo, "A Case Study of Geoacoustic Inversion Based on Convolutional Neural Network Using Vertical Array Data," in *5th International Conference on Information Communication and Signal Processing (ICICSP)*, (Shenzhen, China), pp. 1–6, IEEE, Nov. 2022.

[18] M. J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M. A. Roch, S. Gannot, and C.-A. Deledalle, "Machine learning in acoustics: Theory and applications," *The Journal of the Acoustical Society of America*, vol. 146, pp. 3590–3628, Nov. 2019.

[19] M. D. Collins, W. A. Kuperman, and H. Schmidt, "Nonlinear inversion for ocean-bottom properties," *J. Acoust. Soc. Am.*, vol. 92, pp. 2770–2783, Nov. 1992.

[20] M. D. Collins, WA. Kuperman, and W. L. Siegmann, "Propagation and inversion in complex ocean environments," in *Full Field Inversion Methods in Ocean and Seismo-Acoustics*, pp. 15–20, The Netherlands: Kluwer Academic Publishers, 1995.

[21] M. D. Collins and L. Fishman, "Efficient navigation of parameter landscapes," *J. Acoust. Soc. Am.*, vol. 98, no. 3, pp. 1637–1644, 1995.

[22] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, "Taking the Human Out of the Loop: A Review of Bayesian Optimization," *Proc. IEEE*, vol. 104, pp. 148–175, Jan. 2016.

[23] S. Greenhill, S. Rana, S. Gupta, P. Vellanki, and S. Venkatesh, "Bayesian Optimization for Adaptive Experimental Design: A Review," *IEEE Access*, vol. 8, pp. 13937–13948, 2020.

[24] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press, 2006.

[25] D. R. Jones, "A Taxonomy of Global Optimization Methods Based on Response Surfaces," *J. Global Optim.*, vol. 21, no. 4, pp. 345–383, 2001.

[26] W. F. Jenkins II, P. Gerstoft, and Y. Park, "Bayesian optimization with Gaussian process surrogate model for source localization," *J Acoust. Soc. Am.*, vol. 154, pp. 1459–1470, Sept. 2023.

[27] M. B. Porter, "The KRAKEN normal mode program," *SACLANT Undersea Research Centre Memorandum (SM-245)/Naval Research Laboratory Memorandum Report 6920*, Sept. 1991.

[28] P. Gerstoft and C. F. Mecklenbräuker, "Ocean acoustic inversion with estimation of *a posteriori* probability distributions," *J. Acoust. Soc. Am.*, vol. 104, pp. 808–819, Aug. 1998.

[29] C. F. Mecklenbräuker and P. Gerstoft, "Objective functions for ocean acoustic inversion derived by likelihood methods," *Journal of Computational Acoustics*, vol. 8, no. 2, pp. 259–270, 2000.

[30] K. P. Murphy, *Probabilistic Machine Learning: An Introduction*. Cambridge, MA: MIT Press, 2022.

[31] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *Int. Conf. Learning Representations*, 2019.

[32] N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger, "Information-Theoretic Regret Bounds for Gaussian Process Optimization in the Bandit Setting," *IEEE Trans. Inf. Theory*, vol. 58, no. 5, pp. 3250–3265, 2012.

[33] I. M. Sobol', "On the distribution of points in a cube and the approximate evaluation of integrals," *USSR Comp. Math. and Math. Phys.*, vol. 7, pp. 86–112, Jan. 1967.

[34] A. B. Owen, "Scrambling Sobol' and Niederreiter–Xing points," *Journal of Complexity*, vol. 14, no. 4, pp. 466–489, 1998.

[35] M. Balandat, B. Karrer, D. R. Jiang, S. Daulton, B. Letham, A. G. Wilson, and E. Bakshy, "BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization," in *Advances in Neural Information Processing Systems*, vol. 34, 2020.

[36] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A Limited Memory Algorithm for Bound Constrained Optimization," *SIAM J. Sci. Comput.*, vol. 16, no. 5, pp. 1190–1208, 1995.

[37] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimization," *ACM Trans. Math. Software*, vol. 23, pp. 550–560, Dec. 1997.

[38] Marine Physical Laboratory, "SWellEx-96 Experiment." http://swellex96.ucsd.edu/.

[39] G. L. D'Spain, J. J. Murray, and W. S. Hodgkiss, "Mirages in shallow water matched field processing," *J. Acoust. Soc. Am.*, vol. 105, pp. 3245–3265, June 1999.

[40] R. T. Bachman, P. W. Schey, N. O. Booth, and F. J. Ryan, "Geoacoustic databases for matched-field processing: Preliminary results in shallow water off San Diego, California," *J. Acoust. Soc. Am.*, vol. 99, pp. 2077–2085, Apr. 1996.

[41] D. Eriksson and M. Jankowiak, "High-Dimensional Bayesian Optimization with Sparse Axis-Aligned Subspaces," in *Proc. Mach. Learn. Res.*, vol. 161, pp. 493–501, 2021.

[42] L. Papenmeier, L. Nardi, and M. Poloczek, "Increasing the Scope as You Learn: Adaptive Bayesian Optimization in Nested Subspaces," in *Advances in Neural Information Processing Systems*, vol. 36, 2022.

Chapter 7

Conclusion

This dissertation investigated using unsupervised machine learning for exploratory analysis of continuously recorded seismic data. A key step in the method was the data-driven dimensionality reduction of spectrograms using a convolutional autoencoder, which improved the clustering performance of seismic signals. Automatic identification of dominant forms of seismicity within the data set enabled a more detailed examination of how certain environmental parameters may be associated with specific types of seismicity on the Ross Ice Shelf, as the clustering analysis provided information on *what kinds* of signals are detected, in addition to when and where. The unsupervised aspect of this method can provide time and computational savings, as analysts can use clustering results as an entry point for more targeted interrogations of the data rather than performing manual searches. The method is sufficiently flexible to employ any number of signal detection algorithms, autoencoder architectures, and clustering algorithms for the latent space, and is appropriate for both seismic and acoustic data sets.

Deep clustering continues to find use in both seismological and acoustical applications [1, 2, 3, 4, 5]. The method is particularly promising for the extremely large data sets generated by an emerging field of remote sensing called distributed acoustic sensing (DAS), which utilizes fiber optic cables as arrays that can measure ground motion [6]. These systems record at high sampling rates with high spatial resolution over hundreds or thousands of meters, generating tremendous amounts of data. Deep clustering has been demonstrated for DAS data and will be

essential for exploring these large data sets in an efficient manner [7].

This dissertation additionally presented Bayesian optimization (BO) for efficiently performing geoacoustic inversion with underwater acoustic data. BO was first demonstrated for acoustic source localization in a shallow water waveguide using simulated and experimental data. In the first investigation of BO, the parameter space consisted of source depth and range, with all other geoacoustic parameters fixed. BO successfully localized the source within 144 evaluations of the forward model. A subsequent investigation of BO increased the parameter space dimensionality by estimating array tilt in addition to source localization. BO successfully estimated all three parameters in just 64 evaluations of the forward model and achieved superior performance compared to localization alone, since array tilt estimation reduced the mismatch between the real and modeled environments. Finally, a high-dimensional geoacoustic inversion was performed, in which source localization, array tilt, water depth, and sediment layer properties were jointly estimated. BO correctly estimated the most sensitive parameters and yielded plausible estimates for the least sensitive parameters within 500 evaluations of the forward model.

Unlike exhaustive search and Markov chain Monte Carlo (MCMC) methods, BO is capable of exploring parameter spaces without requiring thousands of evaluations of the forward model. Furthermore, the BO framework is flexible and customizable, allowing for multiple opportunities to assign prior knowledge to constrain the optimization and improve performance. Parameter space boundaries can be constrained by a physical understanding of the waveguide (e.g., restricted to range-depth resolution cells [8]), by geophysical surveys of seabed properties, and by oceanographic knowledge of the structure of sound speed profiles in particular environments. Such knowledge can be encoded in more sophisticated ways than just parameter space boundaries: covariance function behavior can be controlled by selecting kernel functions that appropriately reflect the expected data, and hyperparameter priors can be assigned based on knowledge of the waveguide and environment. With priors carefully assigned in this manner, the Gaussian process interpolation of the matched field objective function gains physical salience upon which acquisition functions capitalize.

Despite the encouraging performance of BO, there is no perfect global optimization algorithm, and BO can suffer from getting stuck in local optima, especially in cases where improper hyperparameter priors—including the choice of kernel function—are used. Improving BO remains a robust and active area of research, with recent advances centered around parallelization of the acquisition function optimization [9, 10, 11, 12, 13, 14, 15, 16], multi-objective optimization [17, 18, 19, 20], and high-dimensional optimization [21, 22]. The high-dimensional optimization approaches of [21] and [22] are particularly interesting for geoacoustic inversion, as they are premised on the assumption that not all parameters in the parameter space are equally important. This may have useful applications for geoacoustic inversion, where parameters like source-range geometry exhibit strong sensitivity, while others, such as certain sediment properties, can exhibit very little. Automatically adjusting kernel function priors to systematically probe and optimize the most sensitive parameter subspaces could make high-dimensional problems with tens of parameters efficient and feasible [22].

Chapter 3 did not implement the methods discussed above, but instead relied on manual interpretation of data collected under sea ice in the Arctic Ocean. The acoustic analysis presented in this dissertation represents not just the types of data that can be collected from challenging environments using acoustic sensors, but also the opportunities available to apply automated machine learning tasks to acoustic data processing and analysis. For example, data from the acoustic sensors were rich with biological activity, and classification, clustering, and other event detection schemes could be applied to this kind of data to identify common biological calls or anthropogenic activity [23, 24, 1, 25].

Each chapter of this dissertation relied on data collected in real environments, from Antarctica to the Arctic to the waters off southern California. Unsupervised exploration of large data sets, along with efficient geoacoustic inversion and parameter estimation, constitute a broader effort to rely on data-driven methods to learn and infer properties of the environment through which seismic and acoustic waves propagate. In each setting, environmental processes that affect seismic and acoustic signal generation and propagation were encoded into the data.

While physical models can predict these processes with high degrees of accuracy, inverting these models to estimate the parameters that explain the observed data is computationally difficult, and data sets are reaching sizes that are impossible for humans alone to analyze. The two paradigms of environmental characterization investigated in this dissertation—unsupervised machine learning and Bayesian optimization—provide tools that can supplement, enhance, and accelerate conventional analytical workflows, with the ultimate goal of more rapidly gaining new insights into our earth and ocean systems.

7.1 References

[1] E. Ozanich, A. Thode, P. Gerstoft, L. A. Freeman, and S. Freeman, "Deep embedded clustering of coral reef bioacoustics," *J. Acoust. Soc. Am.*, pp. 2587–2601, 2021.

[2] C.-C. Wang, E.-J. Lee, W.-Y. Liao, P. Chen, R.-J. Rau, G.-W. Lin, and C.-R. Chu, "Cluster analysis of slope hazard seismic recordings based upon unsupervised deep embedded clustering," *Seismological Research Letters*, vol. 94, pp. 1877–1891, Apr. 2023.

[3] S. M. Mousavi and G. C. Beroza, "Deep-learning seismology," *Science*, vol. 377, Aug. 2022.

[4] T. Sawi, B. Holtzman, F. Walter, and J. Paisley, "An Unsupervised Machine-Learning Approach to Understanding Seismicity at an Alpine Glacier," *JGR Earth Surface*, vol. 127, p. e2022JF006909, Dec. 2022.

[5] S. M. Mousavi and G. C. Beroza, "Machine Learning in Earthquake Seismology," *Annu. Rev. Earth Planet. Sci.*, vol. 51, pp. 105–129, May 2023.

[6] A. H. Hartog, *An Introduction to Distributed Optical Fibre Sensors*. Boca Raton, FL: CRC Press, Taylor & Francis Group, 2017.

[7] C.-C. Chien, W. F. Jenkins, P. Gerstoft, M. Zumberge, and R. Mellors, "Automatic classification with an autoencoder of seismic signals on a distributed acoustic sensing cable," *Computers and Geotechnics*, vol. 155, p. 105233, Jan. 2023.

[8] S. Kim, G. F. Edelmann, W. A. Kuperman, W. S. Hodgkiss, H. C. Song, and T. Akal, "Spatial resolution of time-reversal arrays in shallow water," *J. Acoust. Soc. Am.*, vol. 110, pp. 820–829, Aug. 2001.

[9] D. Ginsbourger, R. L. Riche, and L. Carraro, "A Multi-points Criterion for Deterministic Parallel Global Optimization based on Gaussian Processes," *Hyper Articles en Ligne*, 2008.

[10] D. Ginsbourger, R. Le Riche, and L. Carraro, "Kriging Is Well-Suited to Parallelize Optimization," in *Computational Intelligence in Expensive Optimization Problems*, vol. 2, pp. 131–162, Berlin, Heidelberg: Springer, 2010.

[11] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian Optimization of Machine Learning Algorithms," in *Advances in Neural Information Processing Systems*, vol. 26, (Lake Tahoe, Calif.), 2012.

[12] J. T. Wilson, R. Moriconi, F. Hutter, and M. P. Deisenroth, "The reparameterization trick for acquisition functions," in *NIPS Workshop on Bayesian Optimization*, (Long Beach, Calif.), Dec. 2017.

[13] P. I. Frazier, "A Tutorial on Bayesian Optimization," July 2018.

[14] J. T. Wilson, F. Hutter, and M. P. Deisenroth, "Maximizing acquisition functions for Bayesian optimization," in *Advances in Neural Information Processing Systems*, vol. 32, 2018.

[15] J. Wang, S. C. Clark, E. Liu, and P. I. Frazier, "Parallel Bayesian Global Optimization of Expensive Functions," *Oper. Res.*, vol. 68, pp. 1850–1865, Nov. 2020.

[16] M. Balandat, B. Karrer, D. R. Jiang, S. Daulton, B. Letham, A. G. Wilson, and E. Bakshy, "BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization," in *Advances in Neural Information Processing Systems*, vol. 34, 2020.

[17] S. Daulton, M. Balandat, and E. Bakshy, "Differentiable Expected Hypervolume Improvement for Parallel Multi-Objective Bayesian Optimization," in *Advances in Neural Information Processing Systems*, vol. 34, 2020.

[18] S. Daulton, M. Balandat, and E. Bakshy, "Parallel Bayesian Optimization of Multiple Noisy Objectives with Expected Hypervolume Improvement," in *Advances in Neural Information Processing Systems*, vol. 35, 2021.

[19] B. Tu, A. Gandy, N. Kantas, and B. Shafei, "Joint Entropy Search for Multi-Objective Bayesian Optimization," in *Advances in Neural Information Processing Systems*, vol. 36, 2022.

[20] S. Daulton, S. Cakmak, M. Balandat, M. A. Osborne, E. Zhou, and E. Bakshy, "Robust Multi-Objective Bayesian Optimization Under Input Noise," in *Proc. Mach. Learn. Res.*, vol. 162, pp. 4831–4866, 2022.

[21] D. Eriksson and M. Jankowiak, "High-Dimensional Bayesian Optimization with Sparse Axis-Aligned Subspaces," in *Proc. Mach. Learn. Res.*, vol. 161, pp. 493–501, 2021.

[22] L. Papenmeier, L. Nardi, and M. Poloczek, "Increasing the Scope as You Learn: Adaptive Bayesian Optimization in Nested Subspaces," in *Advances in Neural Information Processing Systems*, vol. 36, 2022.

[23] M. J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M. A. Roch, S. Gannot, and C.-A. Deledalle, "Machine learning in acoustics: Theory and applications," *The Journal of the Acoustical Society of America*, vol. 146, pp. 3590–3628, Nov. 2019.

[24] A. M. Usman, O. O. Ogundile, and D. J. J. Versfeld, "Review of Automatic Detection and Classification Techniques for Cetacean Vocalization," *IEEE Access*, vol. 8, pp. 105181–105206, 2020.

[25] M. Ibrahim, J. D. Sagers, M. S. Ballard, M. Le, and V. Koutsomitopoulos, "Evaluating machine learning architectures for sound event detection for signals with variable signal-to-noise-ratios in the Beaufort Sea," *J. Acoust. Soc. Am.*, vol. 154, pp. 2689–2707, Oct. 2023.