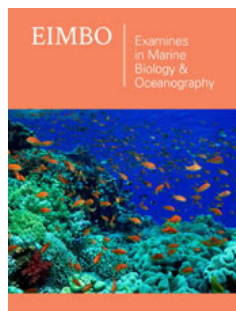



# SeaWatchAI: A Blue/AI Enhancement Illustrating Hybrid Generative AI for Ocean Data Processing, Prediction, and Exploration by Non-Experts at Scale

ISSN: 2578-031X



**\*Corresponding author:** Jason R McKenna, Roger F. Wicker Center for Ocean Enterprise, The University of Southern Mississippi, 1030 30th Ave Gulfport, MS 39501, USA

**Submission:**  June 03, 2024

**Published:**  June 24, 2024

Volume 7 - Issue 1

**How to cite this article:** Jason R McKenna\*, Vishwa Sunkara, James A Thompson, Steve Stanic, Landry Bernard, Craig Harper and Joe Wheeler. SeaWatchAI: A Blue/AI Enhancement Illustrating Hybrid Generative AI for Ocean Data Processing, Prediction, and Exploration by Non-Experts at Scale. *Examines Mar Biol Oceanogr.* 7(1). EIMBO. 000655. 2024. DOI: [10.31031/EIMBO.2024.07.000655](https://doi.org/10.31031/EIMBO.2024.07.000655)

**Copyright@** Jason R McKenna, This article is distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits unrestricted use and redistribution provided that the original author and source are credited.

Jason R McKenna<sup>1\*</sup>, Vishwa Sunkara<sup>1</sup>, James A Thompson<sup>1</sup>, Steve Stanic<sup>1</sup>, Landry Bernard<sup>1</sup>, Craig Harper<sup>2</sup> and Joe Wheeler<sup>2</sup>

<sup>1</sup>Roger F. Wicker Center for Ocean Enterprise, The University of Southern Mississippi, Gulfport, MS 39501, USA

<sup>2</sup>Bluemvmt, Inc., One Broadway, Cambridge, MA 02142, USA

## Abstract

SeaWatchAI revolutionizes ocean data management by offering a proactive approach to environmental monitoring and data management for experts and non-experts at scale. By integrating advanced AI capabilities provided by the Bluemvmt platform with a user-friendly front-end, SeaWatchAI addresses the pressing challenges in hydrography, acoustic data processing, and Uncrewed Surface Vehicle (USV) operations. The platform focuses on three main areas: Collecting, processing, and validating data; synthesizing different types of data from disparate sources, leading to new understandings of the oceanographic environment; and data discovery, accessing, manipulation, storage, and dissemination. By leveraging hybrid generative AI, SeaWatchAI assists with edge processing, prediction, and uncrewed vessel data acquisitions/operations, providing a scalable and adaptable solution for various maritime activities. This alignment with strategic goals helps develop and deploy innovative, cost-effective solutions for managing data collection and addressing environmental challenges.

We present a case study to explore the application of generative AI in enhancing the quality of acoustic Bottom Scattering Loss (BSL) data by allowing the AI to select the best approach to filling data gaps due to random drop-outs, and correcting errors based on outlier analysis. Traditional user-driven methods for BSL QA/QC are time-consuming, require significant human and material resources, and often suffer from data fragmentation. By leveraging generative AI-assisted data correction and imputation, error corrections using Isolation Forest analysis and gap filling using multiple imputation methods are demonstrated. The AI-corrected data shows a closer alignment with the true scattering loss, reducing the mean squared error and improving overall data integrity.

**Keywords:** Hybrid generative; Ocean data; AI; Uncrewed surface; Maritime activities; Isolation forest analysis; SeaWatchAI

## Introduction

SeaWatchAI is poised to revolutionize ocean data management through cutting-edge AI technologies, offering a proactive approach to environmental monitoring and data management. Oceanographic Data Management includes:

- a. Collecting, processing, and validating data;
- b. Synthesis-integrating different types of data from disparate sources, both structured and unstructured data types, leading to “new” understanding of the oceanographic environment; and

c. Data discovery, accessing, manipulation, storage, and dissemination. The emerging potential of Hybrid Generative AI to intelligently assist with edge processing, prediction, and Uncrewed Vessel data acquisitions/operations provides a scalable and adaptable solution for various maritime activities. SeaWatchAI aligns with strategic goals to develop and deploy innovative, cost-effective solutions for managing data collection and addressing environmental challenges, especially enabling the data synthesis functioning capability.

The complexities of ocean data acquisition and analysis are well-documented, often hindered by data fragmentation, accessibility issues, and the need for specialized skills. Inspired by systems like Bluemvmt's Blue/AI data warehouse and CUBEnet [1], USM's data aggregation platform, SeaWatchAI leverages a Retrieval Augmented Generator (RAG) framework and robust data orchestration technologies to streamline these challenges. This article delves into the innovative applications of SeaWatchAI in real-world scenarios to enhance data quality and decision-making processes.

## Generative AI and Retrieval Augmented Generators (RAGs)

Generative AI is a class of artificial intelligence designed to create new content, from text to images, by learning from large datasets. In the realm of oceanic data processing and management, hybrid generative AI systems, particularly those employing Retrieval Augmented Generators (RAGs), are exceptionally promising. RAGs combine the capabilities of generative models with retrieval-based models, enabling them to generate information that is not only contextually relevant but also firmly grounded in real data. This technology can dynamically draw from an extensive corpus of oceanographic data to enhance predictions and outputs, making it especially valuable for various maritime operations at the tactical edge [2-6].

## SeaWatchAI Dashboard

The SeaWatchAI Dashboard is the central hub for accessing and interpreting the predictive analytics provided by SeaWatchAI. It is specifically designed to aid in the acquisition, Quality Assurance and Quality Control (QA/QC), management, and prediction of data from both crewed and uncrewed operations at the tactical edge. Key features of version 1.0 include:

### Navigation bar

Provides links to home, data upload, live map, settings, and help.

### Home page

**Overview panel:** Displays the current operational risk level in user-defined regions, with color-coded indicators for risk levels (Low, Medium, High).

**Real-Time data feed:** Shows the latest data from buoys, satellites, uncrewed surface vehicles (USVs), and crowd-sourced inputs.

**Data prediction panel:** Lists upcoming predictions with confidence intervals and relevant environmental factors for selected operations.

### Data upload portal

Allows users to upload CSV, JSON, GeoJSON, PDFs and enter crowd-sourced observations. Direct ingestion of API data sets is performed via Sidecars.

### Live map

**Interactive map:** Users can select specific regions (e.g., north-central Gulf of Mexico) to view detailed environmental conditions and AI forecasts for uncrewed and crewed operations.

**Filter options:** Allows users to display data layers for temperature, salinity, chlorophyll, and more.

### Settings and custom alerts

Users can customize alerts based on specific or dynamic thresholds for environmental conditions or data QA/QC standards.

Figure 1 is a series of screenshots of the SeaWatchAI environment illustrating a typical session. The top image is the environment after creating a data card for the scattering \*.csv file writing the message to the AI "explore my data". Note that it is also possible to upload PDFs to augment the AI as well. The middle image is the AI being told to "perform all" of the suggested analyses and beginning to generate python code to perform the request. The third image illustrates the final output and SeaWatchAI asking for follow-up instructions. The final image shows that users can save their interactions with the AI (in this case "Analysis Results") to resume the interactions in the main environment at a later date, or potentially show only the python code or images generated up to the save point. The "Saved Results" card can also export the session as a PDF file. For more advanced users, it is also possible to launch a "Sidecar" which could be RStudio, or a Jupyter Notebook for doing advanced analysis or running a dedicated machine learning model generated from data exploration with the SeaWatchAI environment. This sidecar can be read, write, or read/write linked to the main environment for data access or just for passing results. But for most non-expert users, simple commands to load/process, exploit, and disseminate (or "PED") will be sufficient. Regardless of the expertise of the user, the ability to deploy containerized or even specialized versions of the environment guarantees the scalability of the AI even at the tactical edge.

**SeaWatchAI**

Sources: Scattering\_Strength\_Data.csv

Data Insight AI Search

Save

Hello! I am SeaWatchAI, and I've analyzed your dataset, named Scattering\_Strength\_Data.csv, which includes the following columns:

- Grazing Angle (Degrees)**: A continuous variable representing the angle at which the observation is made.
- Frequency (kHz)**: A discrete variable indicating the frequency of the observation in kilohertz.
- Scatter Strength (dB)**: A continuous variable indicating the scatter strength in decibels. Note that there are some missing values in this column.

Here are a few analytical questions you might consider exploring:

- Distribution Analysis**:
  - What is the distribution of Grazing Angles, Frequencies, and Scatter Strengths?
  - Are there any patterns or outliers in these distributions?
- Correlation Analysis**:
  - How does Scatter Strength correlate with Grazing Angle and Frequency?
  - Are there specific frequencies where Scatter Strength shows significant variation with different Grazing Angles?
- Missing Data Analysis**:
  - What is the pattern of missing values in the Scatter Strength column?
  - Can you identify any dependencies or reasons for the missing data?
- Trend Analysis**:
  - How does Scatter Strength trend with Grazing Angle at different Frequencies?
  - Are there specific frequencies where trends in Scatter Strength and Grazing Angle are particularly notable?
- Statistical Analysis**:
  - What statistical significance exists between the Scatter Strength at different Grazing Angles and Frequencies?
  - Can you conduct a hypothesis test (e.g., ANOVA) to determine if there are significant differences in Scatter Strength across different groups?
- Data Visualization**:
  - Plot Scatter Strength against Grazing Angle for each Frequency.
  - Create heatmaps or 3D plots to better visualize the relationship between the three variables.

Would you like to explore any specific aspects or questions in more detail?

1 sources Perform all

Move To Sidebar Remove selected

Saved Results: Your saved results will appear here

**SeaWatchAI**

Sources: Scattering\_Strength\_Data.csv

Data Insight AI Search

Save

- Plot Scatter Strength against Grazing Angle for each Frequency.
- Create heatmaps or 3D plots to better visualize the relationship between the three variables.

Would you like to explore any specific aspects or questions in more detail?

You Perform all

Generated Code Copy Code

```

import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np

# Set display options for inline visualizations in notebooks
%matplotlib inline

# Distribution Analysis
def plot_distributions(df):
    # Plot distributions for each variable
    fig, axs = plt.subplots(3, 1, figsize=(12, 18))

    sns.histplot(df['Grazing Angle (Degrees)'], bins=30, kde=True, ax=axs[0])
    axs[0].set_title('Distribution of Grazing Angles')
    axs[0].set_xlabel('Grazing Angle (Degrees)')
    axs[0].set_ylabel('Frequency')

    sns.histplot(df['Frequency (kHz)'], bins=len(df['Frequency (kHz)'].unique()), kde=True, ax=axs[1])
    axs[1].set_title('Distribution of Frequencies')
    axs[1].set_xlabel('Frequency (kHz)')
    axs[1].set_ylabel('Frequency')

    sns.histplot(df['Scatter Strength (dB)'], bins=30, kde=True, ax=axs[2])
    axs[2].set_title('Distribution of Scatter Strength (dB)')
    
```

1 sources Start typing, Shift + Enter for a new line

Move To Sidebar Remove selected

Saved Results: Your saved results will appear here

**SeaWatchAI**

Sources: Scattering\_Strength\_Data.csv

Data Insight AI Search

Save

**Trend Analysis**

**Scatter Strength vs. Grazing Angle at Different Frequencies:**

- Scatter plots for each frequency provide insight into how Scatter Strength varies with Grazing Angle. There are different trends observable at different frequencies. Scatter Plots

**Statistical Analysis**

**ANOVA Results:**

Source	Sum of Squares	df	F-value	P-value
Frequency (kHz)	135.76	2	1.592	0.205
Grazing Angle (Degrees)	12897.72	1	302.487	2.59e-45
Residual	18958.23	257	NaN	NaN

- Frequency (kHz)**: The F-value is 1.592 (p-value = 0.205), suggesting that the frequency does not significantly impact Scatter Strength.
- Grazing Angle (Degrees)**: The F-value is 302.487 (p-value = 2.59e-45), indicating a highly significant effect of Grazing Angle on Scatter Strength.

**Summary**

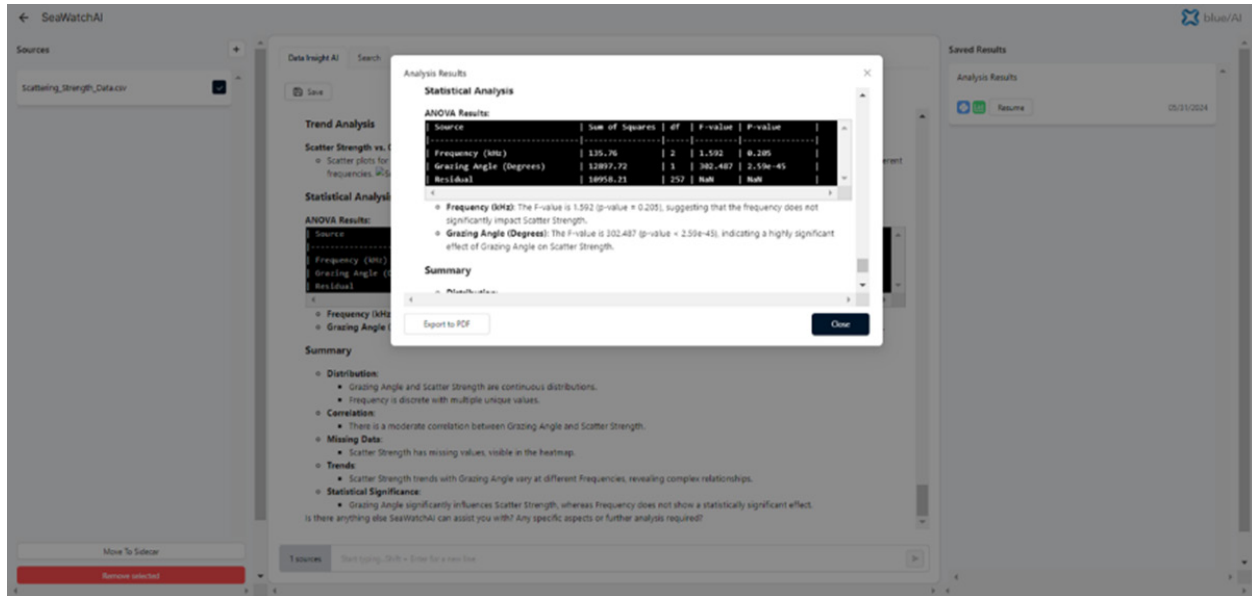
- Distribution**:
  - Grazing Angle and Scatter Strength are continuous distributions.
  - Frequency is discrete with multiple unique values.
- Correlation**:
  - There is a moderate correlation between Grazing Angle and Scatter Strength.
- Missing Data**:
  - Scatter Strength has missing values, visible in the heatmap.
- Trends**:
  - Scatter Strength trends with Grazing Angle vary at different frequencies, revealing complex relationships.
- Statistical Significance**:
  - Grazing Angle significantly influences Scatter Strength, whereas Frequency does not show a statistically significant effect.

Is there anything else SeaWatchAI can assist you with? Any specific aspects or further analysis required?

1 sources Start typing, Shift + Enter for a new line

Move To Sidebar Remove selected

Saved Results: Analysis Results, Resume, 05/31/2024



**Figure 1:** SeaWatchAI data insight environment illustrating the progression of a typical session.

## Features and Benefits

### Data aggregation and accessibility

SeaWatchAI is built upon the Blue/AI product and aggregates vast amounts of public and private oceanic data, making it easily accessible through an AI-enhanced, intuitive search engine that integrates seamlessly with SeaWatchAI.

### Automated data management

The Bluemvmt platform's Blue/FLOW orchestration layer automates the cleaning and normalization of datasets, enhancing data quality and usability.

### Advanced analytical capabilities

Users can connect custom analytical tools and models, such as Jupyter Notebook or Apache Superset, through sidecar technology, fostering flexibility and innovation.

### Data versioning

Provides users with the ability to track data lineage, copyright, and governance issues.

### Collaboration and sharing functionality

Allows users from different labs to collaborate at the object level and share datasets, results, or sidecars with ease.

## Case Study: Acoustic Bottom Scattering Loss Data Processing

Bottom Scattering Loss (BSL) is a critical parameter that affects sonar performance, underwater communication, and naval operations [7,8]. Accurate BSL measurements are essential for understanding the acoustic bottom environment and optimizing sonar systems. However, traditional data collection methods often encounter errors and gaps due to various factors such as equipment malfunction, environmental conditions, and human error. This

case study explores the application of generative AI to enhance the quality of BSL data by filling data gaps and correcting errors. Applied generative AIs deployed on ships such as SeaWatchAI could augment the traditional methods used for executing bottom loss surveys by enhancing these surveys using advanced AI technologies.

### Traditional navy methods for bottom scattering loss surveys use several types of survey and techniques for data collection:

**Shipborne surveys:** Traditional surveys are conducted using ships equipped with sonar systems. These ships traverse designated areas, transmitting acoustic signals and recording the scattered waves.

**Sonar buoys:** Deployed sonar buoys that transmit, receive, and record the bottom scattered acoustic signals at various depths.

### Traditional bottom loss data is a mixture of bottom signal scattering loss and manual QA/QC

**Digital analysis:** Shipboard computers analyze the scattered acoustic signals to determine the scattering loss. This involves filtering noise and calculating the strength of the reflected signals from different bottom types.

**Manual QA/QC:** Shipboard experts manually review the data to ensure accuracy and filter out incorrect or low-quality data.

### Equally important considerations are environmental factors such as

**Water properties:** Seawater properties such as temperature, salinity, and pressure, all of which affect sound speed and transmission loss.

**Seafloor composition:** A priori characterization of the seabed (sand, mud, bedrock, volcanics, etc.) using core samples and historical data.



The challenges of deployed personnel producing timely and relevant data products and information are significant. Acquisition of bottom loss data is time consuming. Manual data collection and processing can be slow, often taking weeks to months for comprehensive surveys. It requires significant human and material resources, including ship time and specialized personnel with specialized training. A final consideration is that these types of surveys suffer from data fragmentation. Often, data from different sources and different acquisition periods may lack consistency, affecting the reliability of the overall survey results. For example, data collection interrupted by severe weather conditions may require a complete re-acquisition if existing stratification and bottom sediments are disturbed sufficiently (especially in shallower regions). SeaWatchAI leverages hybrid generative AI and edge computing to streamline these bottom loss surveys, making them more efficient, accurate, and accessible to minimize these challenges. Using AI can rapidly integrate data from various sensors, perform the necessary data cleaning and imputation, apply real-time filtering and advanced processing, utilize embedded machine learning algorithms to enhance data quality and predictive capabilities while preserving provenance, and ultimately provide quantitative confidence intervals [5,6]. All for non-expert users who are trained in prompt engineering but minimally trained in acoustics and signal processing. Some of the advantages generative AI can introduce to these types of surveys are:

- a. Integrated Data Collection from uncrewed surface vehicles (USVs) and uncrewed underwater vehicles (UUVs) equipped with advanced sonar systems to continuously collect data alongside crewed vessels. Vintage, or even crowd-sourced data can also be seamlessly incorporated (e.g., data from commercial vessels and scientific expeditions) to expand the datasets and can be reprocessed when convenient.
- b. Tactical Edge Computing to process data on-board USVs/UUVs using COTS edge computing devices can enable real-time analysis and decision-making. AI-Driven QA/QC that uses machine learning algorithms to automatically perform quality assurance and control can now identify and correct errors instantly while on-site, dramatically improving mission effectiveness.
- c. Enhanced Predictive Models from hybrid generative AI which combines generative AI with retrieval-based models to generate contextually relevant and accurate predictions of bottom scattering loss. This capability is even more relevant when environmental? The integration of real-time environmental data (temperature, salinity, pressure) to refine predictions can dramatically improve derived data products by highlighting potentially problematic data before surveys are completed.
- d. Blockchain Integration can enable secure data logging so that data is cryptologically verifiable, and chain of custody is established for various transactions ensuring traceability and immutability. This will allow continuous model updates using deployed personnel's experiences in highly disadvantageous situations. This process of continuous updates to the AI using

new, secure data and human engagements will enhance their predictive accuracy over time.

## Generative AI-Assisted Data Correction & Imputation

In this example, SeaWatchAI leverages hybrid generative AI and edge computing to streamline BSL surveys, making them more efficient, accurate, and accessible. The platform integrates data from various sensors, applies real-time processing, and utilizes machine learning algorithms to enhance data quality and predictive capabilities. This data can include errors and gaps, which represent common issues encountered during data collection. Below, we demonstrate the application of generative AI to illustrate potential improvements in data quality through the following automated processes:

### Error correction

Isolation Forest analysis can automatically identify outliers in the data. By marking these anomalies as missing values and subsequently imputing them, the AI can effectively correct the outliers and bring the data closer to the true scattering strength trends.

### Gap filling

Using multiple imputation methods can allow the AI to select the most successful approach to filling the missing values due to acquisition dropouts. When operationalizing the AI to assist with errors and data dropouts, chaining the processes together to remove outliers prior to imputation would be the preferred route. However, for the purpose of maximizing clarity, we discuss each approach separately.

### Isolation forest analysis and visualization

A typical acoustic scattering file contains the following columns (at a minimum):

#### Grazing Angle (Degrees)

Continuous values representing the grazing angle for the observed scatter strength.

#### Frequency (kHz)

Continuous values representing the frequency of the scattering strength in kHz.

#### Scatter Strength(dB)

Continuous values representing the scattering strength with some missing values.

As a first step, the AI first identified the missing values in the Scatter Strength column. It then selected the Isolation Forest method [9] to detect anomalies. Isolation Forest is an unsupervised learning algorithm specifically designed for anomaly detection. It uses the concept of isolating observations to identify anomalies. In this context "Isolation" means anomalies are few and different. Thus, they are easier to isolate than normal points. Additionally, "Isolation Trees" or "iTree" are binary trees used to recursively

divide the data. The length from the root to a leaf node helps determine if a point is an anomaly. The algorithm works as follows: each data point is evaluated for its degree of anomaly, and points are classified as outliers or normal based on their isolation scores:

### Random subsampling

The data is subsampled to create  $n$  different data subsets.

### Tree construction

For each subset, an iTTree is constructed by:

- Randomly selecting a feature.
- Randomly selecting a split value between the minimum and maximum values of the selected feature.

### Recursive splitting

The splitting process continues until:

- The data point is isolated (a leaf node).
- The predefined maximum tree height is reached.

### Anomaly scoring

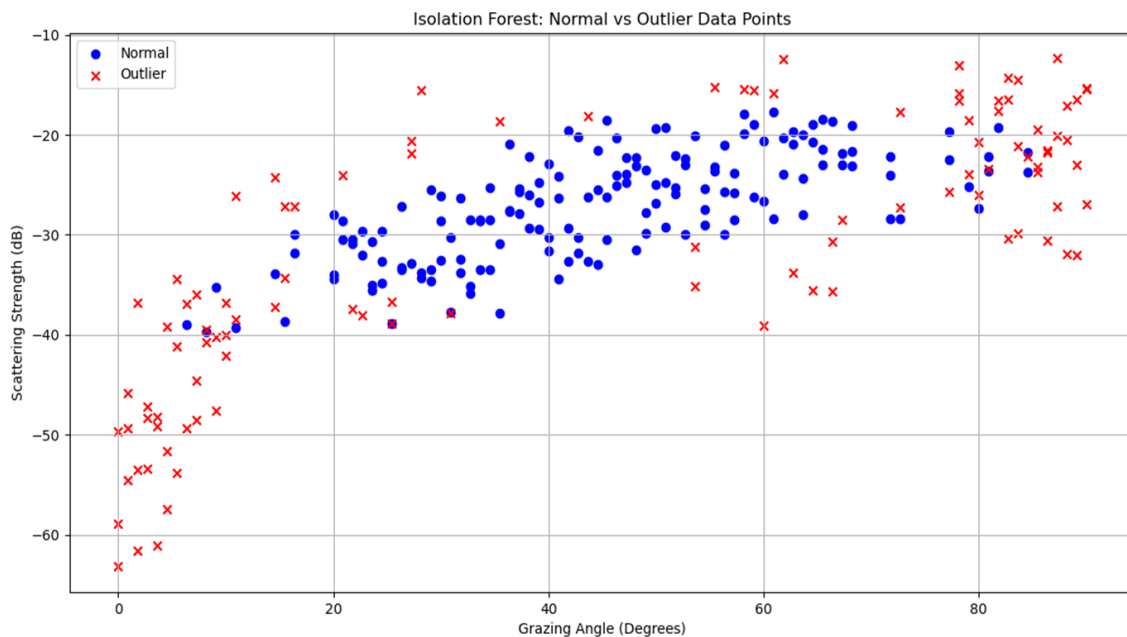
- Average path length is calculated for normal points over  $n$  iTrees.

- The anomaly score is derived from the path length. Points with shorter average path lengths have higher anomaly scores, indicating they are anomalies.

The Isolation Forest approach has several advantages, including:

- Efficiency:** It handles large datasets efficiently.
- Adversarial robustness:** There is less sensitivity to swamping and masking effects.
- Interpretability:** Models are easy to understand and visualize.

To determine if the AI-identified anomalies are valid outliers, we can visualize the data distributions, perform a statistical analysis of the anomalies compared to the rest of the data and perform a contextual investigation to understand the domain to provide a contextual analysis of the anomalies. To assist in visualizing the entire dataset, in Figure 2 we present the distribution of features and identified anomalies. This helps to understand the context in which anomalies appear. Here, the AI selected outliers as red triangles vs blue dots to distinguish between normal and anomalous data points. The key findings from the analysis are discussed below.



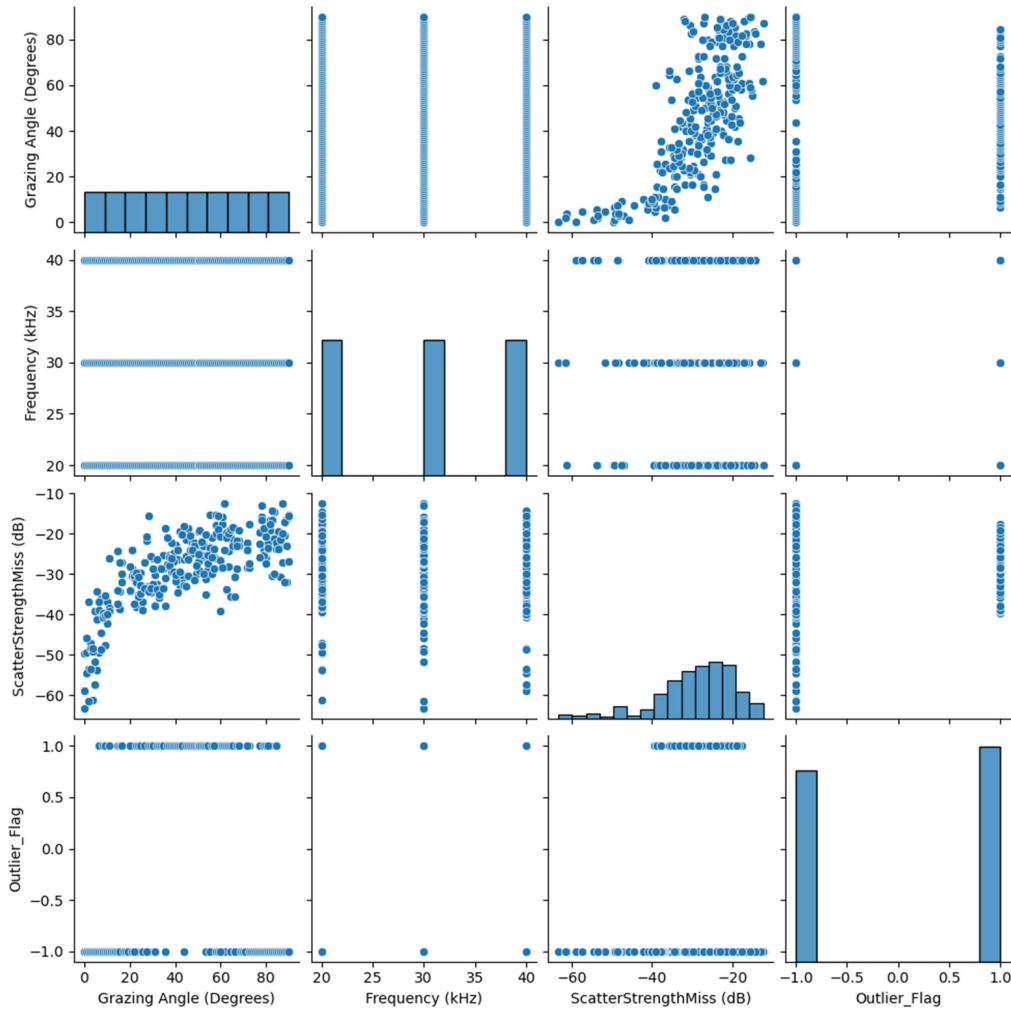
**Figure 2:** Isolation forest analysis visualization. The isolation forest was fitted on the entire dataset (with missing values filled).

In addition to visually inspecting the AI selected outliers, we can use pair plots to statistically represent the data (Figure 3) to provide a comprehensive visualization of the relationships between Grazing Angle (Degrees), Frequency (kHz), and Scatter Strength (dB). In general, pair plots are a valuable tool

for exploratory data analysis, offering insights into potential correlations and distributions among variables. When the data was preprocessed and missing values were imputed, the pair plot analysis illustrates the underlying patterns that could be indicative of the data's structure and relationships. Here we can hypothesize

that clear relationships between the variables are discernible. The diagonal elements show the distribution for each variable, with Grazing Angle and Scatter Strength exhibiting relatively normal distributions. Off-diagonal scatterplots reveal correlations between Grazing Angle and Scatter Strength, indicating that as the grazing angle increases, the scatter strength tends to vary in a recognizable pattern. The Frequency variable, while categorical, shows distinct

clusters in scatter plots, corresponding to the frequency categories present in the data. Notably, several outliers are visible, particularly in the plots involving Scatter Strength, suggesting the presence of anomalous measurements or unique scattering conditions. These observations provide a foundational understanding of the interplay between variables and serve as a precursor to more detailed statistical analyses.



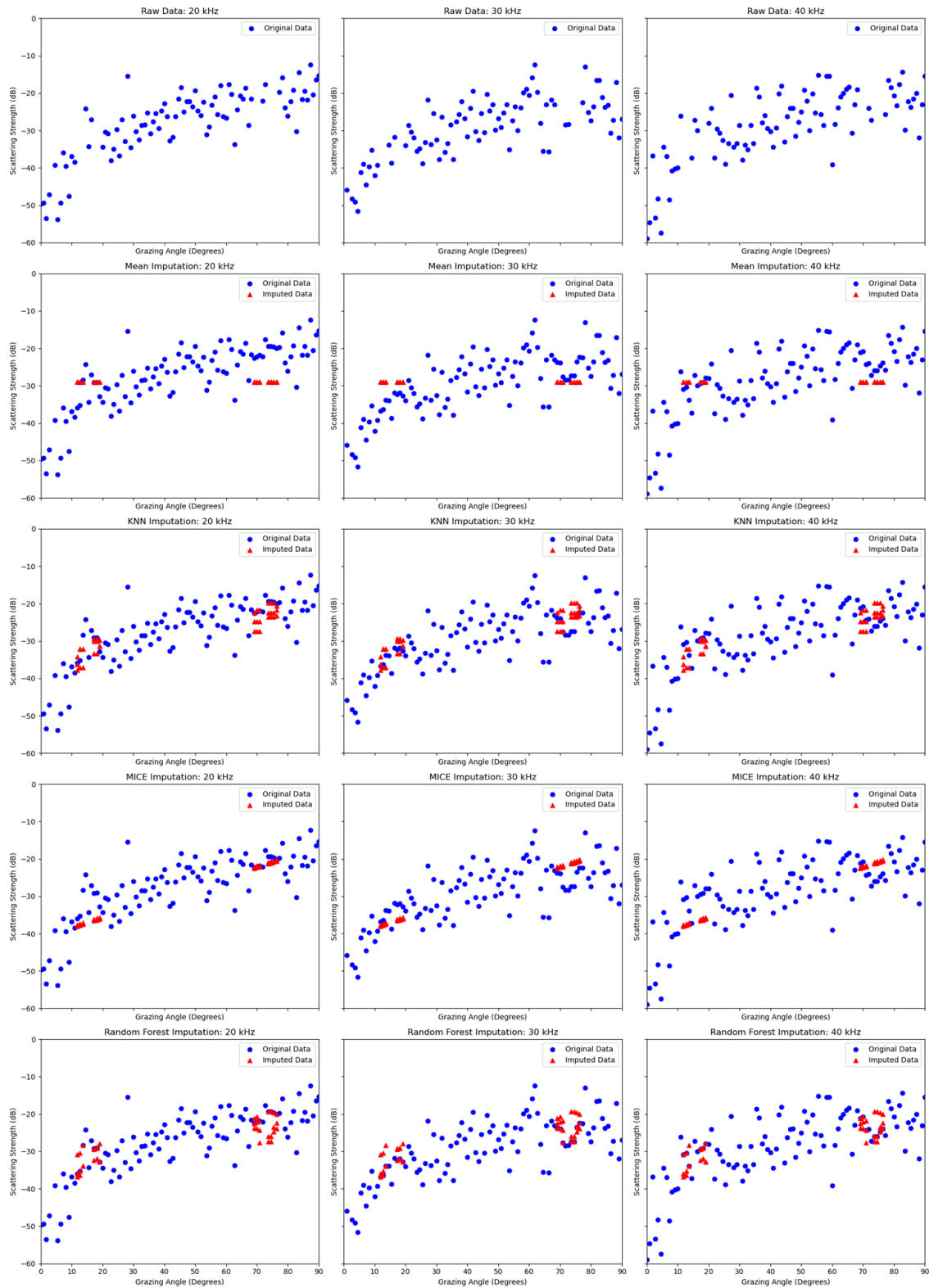
**Figure 3:** Pairplot of the acoustic scattering strength using the raw data and isolation forest analysis.

Contextually, Isolation Forest analysis performed on our dataset of acoustic scattering strength measurements, identified 30 outliers out of 261 entries. These anomalies were detected based on the attributes of Grazing Angle (Degrees), Frequency (kHz), and Scatter Strength(dB). The remaining 231 entries were classified as inliers, indicating normal behavior. The detection of these outliers is crucial as they could signify potential measurement errors, unique scattering conditions, or other underlying factors that warrant further investigation. This AI-driven analysis underscores the importance of utilizing robust anomaly detection techniques in the acquisition process to ensure data integrity and to highlight unusual observations that may impact subsequent analyses and interpretations. While the Isolation Forest aided in detecting most outliers effectively, ensuring data integrity for subsequent analyses, it did tend to misclassify non-linear trends as outliers. This can be

overcome with tuning and the addition of checks from temporal and spatial analysis.

### Imputation Methods and Visualization

To better understand scattering loss data with missing values, we applied different imputation methods to the same acoustic scattering strength dataset and analyzed the results. The missing values were found at various grazing angles across frequencies (20kHz, 30kHz, and 40kHz). Figure 4 presents the scattering strength data by frequency. Blue points represent original data, while red triangles indicate the imputed values. Gaps in the data are present at 4 grazing angle ranges shown in Table 1 (39 total drop-outs) and are indicated by missing values. The four methods selected by the AI for imputation analysis are:



**Figure 4:** Scatter strength data dropout imputation using Mean, KNN, MICE, and random forest methods.

**Table 1:** Scatter strength data drop-outs.

Row Number	Grazing Angle (Degrees)	Frequency (kHz)	Scatter Strength (dB)
13	11.82	20	NaN
14	12.73	20	NaN
15	13.64	20	NaN



19	17.27	20	NaN
20	18.18	20	NaN
21	19.09	20	NaN
76	69.09	20	NaN
77	70	20	NaN
78	70.91	20	NaN
81	73.64	20	NaN
82	74.55	20	NaN
83	75.45	20	NaN
84	76.36	20	NaN
113	11.82	30	NaN
114	12.73	30	NaN
115	13.64	30	NaN
119	17.27	30	NaN
120	18.18	30	NaN
121	19.09	30	NaN
176	69.09	30	NaN
177	70	30	NaN
178	70.91	30	NaN
181	73.64	30	NaN
182	74.55	30	NaN
183	75.45	30	NaN
184	76.36	30	NaN
213	11.82	40	NaN
214	12.73	40	NaN
215	13.64	40	NaN
219	17.27	40	NaN
220	18.18	40	NaN
221	19.09	40	NaN
276	69.09	40	NaN
277	70	40	NaN
278	70.91	40	NaN
281	73.64	40	NaN
282	74.55	40	NaN
283	75.45	40	NaN
284	76.36	40	NaN
			Number of Dropouts: 39

a) Mean Imputation replaces missing values with the mean of the respective column. This method assumes that the data is Missing Completely at Random (MCAR). The approach tends to insert a constant mean value, which may not reflect the real data patterns or trends accurately.

b) K-Nearest Neighbors (KNN) Imputation: KNN Imputation fills missing values by finding the nearest neighbors (in our case,  $k=5$ ) and averaging their values. This method leverages the similarity between data points and tends to produce a higher fidelity imputation, closely matching the data trend.

c) Multiple Imputation by Chained Equations (MICE): MICE performs multiple rounds of imputations by predicting missing

values through iterative models. This method can accommodate various kinds of missing data patterns and improves upon single imputation techniques. This method of imputation tends to offer a balanced approach, preserving variability but with slightly higher mean squared error (usually).

d) Random Forest Imputation: uses a Random Forest model to predict and fill in missing values. The model is trained on the available data and then used to predict the missing values.

In all cases the AI was given the data and computed the imputation without user guidance. The upper plot in Figure 4 shows synthetic bottom scattering loss data (in dB) with data dropouts at 4 grazing angles. The data points represent different frequencies

(20kHz, 30kHz, and 40kHz). Often typical gaps introduced by data-dropouts during data collection in ocean acoustics may not be readily apparent. The lower plots show synthetic bottom scattering loss data after applying generative AI-assisted data imputation using the four methods to fill in the gaps caused by data dropouts. The original data points (indicated by blue circles) and the AI-corrected data points (red triangles) show a close alignment with the expected scattering loss, reducing the mean squared error and improving overall data integrity.

### SeaWatchAI Imputation Recommendations

The AI uses the Mean Squared Error (MSE, measures the average squared differences between true and imputed values) and Mean Absolute Error (MAE, measures the average absolute differences between true and imputed values) as the critical metrics for evaluating the performance of above imputation methods. The results of the imputation are summarized in Table 2:

**Table 2:** Summary of errors of imputed data. KNN and random forest result in the lowest errors and also closely align with the scattering strength data trends.

Imputation Method	Mean Squared Error (MSE)	Mean Absolute Error (MAE)
KNN	789.5	27.6
Mean	837.2	28.9
MICE	873.3	28.5
Random Forest	770.7	27.3

- a) Random Forest Imputation yielded the lowest MSE (770.66) and MAE (27.75), indicating that it provided the best approximations compared to the other methods. This underscores the strength of the Random Forest model in capturing complex relationships within the data and efficiently predicting the missing values.
- b) KNN Imputation followed closely, with slightly higher MSE (789.53) and MAE (27.98) compared to Random Forest. This method leverages the local similarity between data points effectively but might struggle with more complex data patterns.
- c) Mean and MICE Imputations resulted in higher error metrics. Mean Imputation’s higher MSE (837.19) and MAE (28.94) suggest it oversimplifies the data by assuming a uniform distribution. MICE Imputation, despite being a more sophisticated method, showed a higher MSE (873.25) and MAE (29.13), potentially due to the inherent complexity or suboptimal performance on this specific dataset.

Even though Random Forest and KNN present clear advantages with respect to minimizing error metrics in the imputation, they have one major disadvantage in that they can be computationally expensive, especially for large datasets, as they require calculating distances between significant numbers of data points or generating many models. Overall, the AI-corrected data shows a closer alignment with the true scattering loss reducing the mean squared error and improving overall data integrity. The combination of KNN imputation and Isolation Forest models ensures robust handling of both missing values and anomalies, making the system adaptable to various oceanographic conditions. The generative AI approach also provides a scalable solution for real-time data correction, significantly reducing the time and resources required for manual QA/QC processes (see Table 3). The improvements in data completeness, accuracy, and reduced error rates underscore the efficacy of SeaWatchAI’s generative AI methods in handling gaps and anomalies in scattering strength data. These advancements provide a robust foundation for enhanced oceanographic surveys and related maritime operations.

**Table 3:** Comparative analysis of processing accuracy with and without generative AI in the hypothetical analysis above.

Metric	Without AI	AI Assisted
Data Processing Accuracy (%)	85.2	95.8
Error Rate (%)	14.8	4.2
Time for Data Processing (hrs)	4	0.2

### AI-Augmented data models to improve operations

One of the more powerful aspects of using Generative AI in workflow is the ability for examining several hypothetical challenges prior to data collection and having the capability to provide solutions at the ready to overcome them. For example:

**Impact of water properties:** Investigate how variations in temperature, salinity, and pressure affect the AI’s performance in correcting BSL data.

**Temporal variability:** Analyze the AI’s effectiveness in handling data collected over different seasons and time frames, or even during rapidly changing environmental conditions. The latter is critically important, considering the dynamic nature of the world’s oceans.

**Model scalability:** Assess the scalability of the AI models when applied to larger datasets and different geographic regions, ensuring consistent performance.

Overall, the integration of generative AI into bottom surveys offers substantial benefits for naval operations and underwater acoustics research. By enhancing data accuracy and efficiency, generative AI such as SeaWatchAI provides a powerful tool for addressing the challenges of traditional data collection methods. This case study illustrates the potential of AI-driven solutions in advancing oceanographic data processing and sets the stage for further research and development in this field.

## Other Potential Oceanographic Missions Impacted by Generative AI

### Hydrography collection at the tactical edge

SeaWatchAI's application in hydrography data collection at the tactical edge highlights its ability to process and verify data in real-time. USVs or Crewed vessels use sophisticated multi-beam echo sounders to capture bathymetric data. This data, often collected in remote or minimally supervised environments, is processed using on-board edge computing. The AI performs instant QA/QC checks to filter out noise and erroneous readings, which is crucial for maintaining the integrity of data used in navigational charts and environmental monitoring. This capability is vital where traditional expert oversight is unfeasible, thus enabling continuous and reliable data collection.

### USV operations control from shore

The AI's role extends to controlling USV operations from shore, particularly under changing weather conditions that could affect data collection, such as in hypoxia mapping on continental shelves. SeaWatchAI integrates real-time meteorological data and oceanic conditions into its operational framework, allowing remote operators to make informed decisions about USV navigation and data collection strategies. This scenario highlights the AI's capacity to adapt to environmental changes and manage operations effectively, ensuring continuity and precision in data collection efforts.

## Conclusion

Generative AI, exemplified by SeaWatchAI, marks a significant advancement in democratizing ocean data exploration and analysis. Built on the robust Blue/AI platform, SeawatchAI leverages cutting edge technologies to lower barriers to entry for non-experts and enhance the speed and accuracy of data-driven decisions, SeaWatchAI aims to become an indispensable tool in global oceanography and environmental science. Overall, the AI made several recommendations to assist in the QA/QC of acoustic scattering strength data that, given the data structure, appear valid

and could aid non-expert users:

- a. Utilizing Isolation Forest for outlier detection can enhance data quality and reliability and result in more robust data drop-out resolution
- b. For datasets with similar properties, Random Forest Imputation is recommended to solve data drop-outs due to its superior performance in addressing missing values.

Integrating SeaWatchAI into these types of projects involving oceanographic data collection and processing at the tactical edge offers substantial advantages over traditional methods. By leveraging advanced AI, edge computing, and micro-blockchain technology, SeaWatchAI, built on the powerful Blue/AI platform, enhances the efficiency, accuracy, and reliability of acoustic bottom scattering loss measurements. This transformation not only enhances naval operational capabilities but also democratizes ocean data processing, making it accessible to a broader range of users and applications. The seamless integration with the Blue/AI platform, CUBEnet, or other data platforms ensures that SeaWatchAI can aggregate, manage, and analyze vast amounts of oceanographic data, providing comprehensive insights and fostering innovation across various maritime activities.

## References

1. Stanic S, Wiggert J, Bernard L, McKenna J, Sunkara V, et al. (2024) Coastal CUBEnet: An integrated observation and modeling system for sustainable Northern Gulf of Mexico Coastal Areas. *Frontiers in Marine Science* 11: 1400511.
2. Radford A, Wu J, Child R, Luan D, Amodei D, et al. (2019) Language models are unsupervised multitask learners. *OpenAI*, pp. 1-24.
3. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, et al. (2020) Language models are few-shot learners. *Computation and Language*, Arxiv Preprint, p. 14165.
4. Chen M, Tworek J, Jun H, Yuan Q, Pinto HP, et al. (2021) Evaluating large language models trained on code. *Machine Learning*, ArXiv Preprint, pp. 1-35.
5. Kar S, McKenna JR, Sunkara V, Coniglione R, Stanic S, et al. (2024) XWaveNet: Enabling uncertainty quantification in short-term ocean wave height forecasts and extreme event prediction. *Appl Ocean Res* 148: 103994.
6. Kar S, McKenna JR, Anglada G, Sunkara V, Coniglione R, et al. (2023) Forecasting vertical profiles of ocean currents from surface characteristics: A multivariate multi-head CNN-LSTM approach. *J Mar Sci Eng* 11(10): 1964.
7. Rubisch JH, Haas PS (1977) Bottom loss measurements in the Eastern Pacific Ocean. Naval Air Development Center Warminster, USA.
8. Urlick RJ (1983) Principles of underwater sound. McGraw-Hill, USA.
9. Liu FT, Ting KM, Zhou ZH (2008) Isolation forest. *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*.