NOVA IMS Information Management School

MGI

Master Degree Program in

Information Management

Enhancing the Quality of ARGO's In Situ Measurements Data Using Machine Learning Algorithms

Vasco Dias Pereira

Project Work

presented as partial requirement for obtaining a Master's Degree in Information Management

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Enhancing the Quality of ARGO's In Situ Measurements Data Using Machine Learning Algorithms

by

Vasco Dias Pereira

Project Work presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Business Intelligence

Supervised by

Márcia Lourenço Baptista, PhD, Nova Information Management School

December, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Oeiras, 05/12/2024

ACKNOWLEDGEMENTS

As I reach the end of this journey, I want to express my heartfelt gratitude to all who supported me along the way.

A special thank you to my supervisor, Márcia Baptista, for her guidance, patience, and insight. Your mentorship was essential to this thesis and my growth.

To my parents, thank you for your constant support and love, and to my sister Madalena, for always lifting me up.

To Marta your encouragement and presence through the toughest times meant everything.

Granny, your Friday lunches brought comfort and strength when I needed them most.

I'm also grateful to my company for allowing me to bridge academic learning with real-world practice.

Lastly, to myself for the dedication, perseverance and hard work. This achievement reflects commitment, resilience, and growth.

Thank you all.

ABSTRACT

The ARGO program is a global initiative that collects in situ oceanographic data through autonomous profiling floats, offering consistent and widespread measurements of temperature, salinity, and pressure across the world's oceans. Despite its broad spatiotemporal coverage, the dataset frequently contains missing values due to sensor malfunctions, data transmission issues, or environmental limitations. These gaps can undermine the reliability of ocean analyses and climate modeling efforts. This thesis explores the use of Self-Organizing Maps (SOM), an unsupervised machine learning technique, for imputing missing values in the ARGO dataset. The methodology follows the CRISP-DM framework, encompassing data transformation, outlier removal, normalization, and simulation of missing values for evaluation. SOM was trained on complete observations using both spatial (latitude and longitude) and oceanographic features. For each incomplete test record, the Best Matching Unit (BMU) on the SOM grid was identified, and nearby neurons were queried to estimate missing values through a distance-weighted averaging strategy based on geographic proximity. Model performance was evaluated by comparing imputed values against artificially masked ground truth using standard metrics. Results show that SOM performed best on salinity (R^2 = 0.54), with moderate accuracy on temperature (R^2 = 0.32) and pressure (R^2 = 0.30). While the SOM model did not outperform the K-Nearest Neighbors (KNN) baseline in terms of error metrics, it demonstrated value in preserving physical coherence and spatial structure. The study highlights SOM's potential for integration into data quality pipelines, especially when interpretability and global pattern recognition are desired.

KEYWORDS

ARGO; Self Organizing Maps; Environment; Data Imputation; Machine Learning for Oceanography

Sustainable Development Goals (SDG):

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AI: Artificial Intelligence
ARGO: Array for Real-time Geostrophic Oceanography
BMU: Best Matching Unit
CRISP-DM: Cross-Industry Standard Process for Data Mining
CSV: Comma-Separated Values
dbar: Decibars
ECMWF: European Centre for Medium-Range Weather Forecasts
FTP: File Transfer Protocol
GDAC: Global Data Assembly Centers
GPS: Global Positioning System
HTML: HyperText Markup Language
KNN: K-Nearest Neighbors
MAE: Mean Absolute Error
mg/m^3: Milligrams per Cubic Meter
MICE: Multivariate Imputation by Chained Equations
ML: Machine Learning
NADW: North Atlantic Deep Water
NOAA: National Oceanic and Atmospheric Administration
pH: Potential of Hydrogen
PSU: Practical Salinity Units
RF: Random Forest
RMSE: Root Mean Squared Error
SOM: Self-Organizing Maps
SVM: Support Vector Machines
W/m^2: Watts per Square Meter
WGAN: Wasserstein Generative Adversarial Network
°C: Degrees Celsius
µmol/kg: Micromoles per Kilogram
µmol/L: Micromoles per Liter

1. INTRODUCTION

The ocean plays a central role in the Earth's climate system. It absorbs the majority of the excess heat generated by greenhouse gas emissions and influences weather patterns, sea level, and marine ecosystems. Understanding its physical properties—such as temperature, salinity, and pressure—is key to monitoring long-term climate change and improving global forecasts.

Since its official launch in 2000, the ARGO program has transformed how ocean data is collected. Thousands of autonomous floats now provide regular vertical profiles across all major oceans. These observations have become a critical data source for climate science and oceanography. Still, challenges persist. Harsh environmental conditions, sensor limitations, and data transmission errors often lead to missing or unreliable values. This affects the quality of the dataset and, in turn, the analyses that depend on it.

Missing data in oceanographic records is not just a technical inconvenience—it can distort scientific conclusions. Addressing this issue calls for imputation techniques that go beyond basic statistical methods. Self-Organizing Maps (SOM), a type of unsupervised learning model, offers a way to fill gaps while preserving the spatial and multivariate structure of the data. They are well-suited to capture complex, nonlinear relationships common in ocean datasets.

This thesis investigates the use of SOMs to impute missing values in the ARGO dataset. The approach follows a structured workflow inspired by the CRISP-DM framework. It covers the collection, transformation, exploration, and preparation of data, followed by model training and evaluation. Artificially removed values are compared with their known ground truth, allowing a quantitative assessment of the method's accuracy. This work contributes to improving the precision of ocean samples and to improve the quality of ARGO data for oceanographic research.

1.1. RESEARCH GAP

The ARGO program has become a cornerstone of global ocean monitoring by providing consistent and large-scale in situ observations of temperature, salinity, and pressure. However, like many observational datasets, ARGO suffers from data quality challenges, particularly missing values, which arise from sensor failures, communication issues, or adverse environmental conditions. These gaps limit the reliability of downstream analyses and hinder the full use of the dataset in modeling ocean dynamics and climate variability.

Traditional methods for handling missing oceanographic data include statistical imputation techniques such as mean substitution or linear interpolation. While these approaches are simple to implement, they often fail to capture the spatial and nonlinear dependencies that

characterize the ocean environment. As a result, they may introduce distortions or bias into the data, especially in cases where the missingness is not random or uniformly distributed.

Recent advances in machine learning offer alternative frameworks that can better model the complex, multivariate structure of oceanographic data. Techniques such as Self-Organizing Maps (SOM) provide a way to preserve topological relationships and spatial patterns during imputation. Although machine learning methods have been applied successfully in other domains for data repair and completion, their use in operational oceanography and specifically in ARGO data imputation remains limited.

This thesis addresses this gap by evaluating the use of SOMs to impute missing values in the ARGO dataset, comparing its performance against traditional statistical methods. The goal is to assess whether machine learning models can improve the accuracy and reliability of imputed data while maintaining the spatial structure and physical coherence of oceanographic profiles. In other terms, to what extent can Self-Organizing Maps outperform traditional imputation methods in reconstructing missing oceanographic data from the ARGO dataset while preserving spatial and physical consistency?

1.2. OBJECTIVES

This thesis addresses several research objectives. The first objective is to understand the data quality issues that are present in the dataset. For example, is it important to investigate if there is only missing data or if there are more data quality issues.

The second objective is to understand the statistical patterns underlying the missing data. For this it is important to understand if there is any pattern in the existing missing data and try to replicate it. For this, questions like how the patterns of missing data can emerge need to be addressed.

The next objective is to understand what machine learning algorithms will be necessary to address and solve the data quality issues. It will be important to analyze which machine learning algorithms should be used and how well they perform against traditional methods like mean statistical imputation.

Finally, the last objective is to evaluate the quality improvement results. The newly calculated values must be assessed for accuracy and reliability. Questions here are how imputation performance can be measured or how well these records need to be imputed in order to be considered viable.

2. LITERATURE REVIEW

This chapter reviews the ARGO project's historical development, technical framework, and addresses challenges related to data quality as well as related work.

2.1. THE ARGO PROGRAM

The ARGO project is a global program initiated in 1999 that monitors ocean water properties using a network of robotic floating instruments. These floats drift with ocean currents and travel vertically between the surface and mid-water depths, having most of their operational life beneath the waves. The program's name, "ARGO," was inspired by its collaboration with the Jason satellite. Interestingly, this name references Greek mythology, when the hero Jason embarked on his legendary quest aboard the ship ARGO in his search for the Golden Fleece.

2.2. HISTORICAL CONTEXT

Having started in 1999 and being officially launched in 2000, the ARGO project consists of a global array of approximately 4,000 autonomous profiling floats with sensor technology that is used to study various components of the ocean, such as temperature, salinity and others (Roemmich et al., 2019). This initiative addresses the limitations of previous shipboard methods, that often resulted in sparse data coverage, particularly in remote regions such as the Southern Ocean (Le Traon et al., 2020). Importantly, the development of the ARGO project and associated datasets follows from the recognition of the ocean's role in the Earth's climate system. Note that over 90% of the heat accumulated over the past 50 years has been absorbed by the oceans (Desbruyères et al., 2016).

The primary purpose of ARGO is to provide continuous, high-resolution measurements of different oceanographic parameters to be used in the analysis of ocean dynamics and their influence on global climate patterns (Roemmich et al., 2019). The data collected by ARGO floats produce monthly global data on temperature, salinity and currents at different depths. This information allows researchers to identify seasonal and interannual variations in ocean properties (Su et al., 2020). Furthermore, the program has facilitated a standardized understanding of the mixed layer's, ocean's surface, and mid-water level characteristics over time and across different regions (mixed layer climatologies) (Sivareddy et al., 2017). Notably, with recent improvements in the technological platform and sensor devices the accuracy of ocean models and forecasting has improved (Roemmich et al., 2019).

ARGO data have become the dominant input for estimating global ocean heat content and associated thermosteric sea level rise (G. C. Johnson et al., 2019). Also, the ARGO program has contributed to an enhanced understanding of the ocean's circulation patterns. For instance, studies utilizing ARGO data have revealed significant changes in deep ocean circulation and warming trends in bottom waters, which have important implications for global climate models (G. C. Johnson, 2022). Also of note is the fact that ARGO has contributed to a better understanding of high-resolution phenomena. For example, it is now better understood the

phenomena of mesoscale eddies and their impact on nutrient distribution and marine ecosystems (Roach et al., 2016).

Oceanic processes such as upwelling and mixing are oceanic movements where deep, cold, and nutrient-rich water rises to the surface, replacing or mixing with the surface water. These processes play a major role in understanding marine productivity and biodiversity, controlling the exchange of properties such as temperature or nutrients (Grégoire et al., 2021). The ARGO floats' ability to measure salinity and temperature at various depths allows for this assessment of ocean mixing and heat distribution (Su et al., 2020). Additionally, the dataset has proven useful in identifying biases in historical oceanographic data, thereby improving the reliability of climate models and predictions (Sivareddy et al., 2017).

Finally, the integration of ARGO measurements with satellite observations has helped to monitor ocean dynamics on a global scale (Sánchez-Román et al., 2017).

With the pilot introduction of Deep ARGO floats the program is extending its coverage to deeper depths beyond 2,000 meters (G. C. Johnson et al., 2019). This expansion is important for understanding deep ocean processes and their role in global climate (G. C. Johnson et al., 2020). The ongoing developments, including the integration of biogeochemical sensors, promise to enhance our understanding of ocean behavior and climatic consequences (Grégoire et al., 2021).

2.3. TECHNOLOGICAL FRAMEWORK

The technological framework of the ARGO project is anchored in the use of autonomous profiling floats, which are equipped with a set of sensors capable of measuring several components:

Component	Unit of Measure
Temperature	ºC
Salinity	Salinity Unit
Pressure	Decibars
Dissolved Oxygen	µmol/kg
pH	pH Units
Nitrate	µmol/L
Chlorophyll-a	mg/m³
Suspended Particles	Arbitrary Units
Downwelling Irradiance	W/m²

Table 1 - Components measured by ARGO float's sensors

These floats are designed to drift with ocean currents, periodically diving to depths of up to 2,000 meters and then ascending to the surface to transmit their data via satellite communication systems. They are programmed to follow a specific cycle, typically spending about 10 days at the surface before descending again to collect additional data. This cycle

allows for continuous monitoring of ocean conditions across vast areas of the world's oceans, which is essential for capturing temporal variations in ocean properties (Wong et al., 2020).

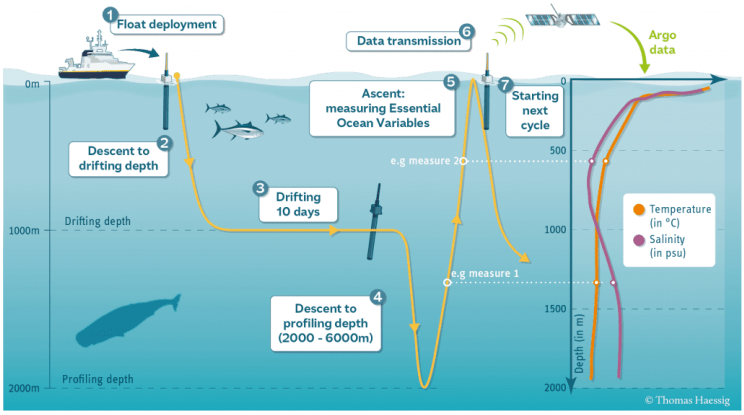


Figure 1 - ARGO Float Cycle (Argo Program, n.d.)

The methodology employed by the ARGO project involves a combination of autonomous data collection and sophisticated data processing techniques. Each float is equipped with a GPS system that allows it to determine its position accurately, which is critical for mapping ocean currents and understanding spatial patterns in ocean temperature and salinity. The data collected by the floats, during the 10 day-cycle, are transmitted to data centers via satellite, enabling near real-time high resolution profile data transmission. There, they undergo quality control and processing to ensure accuracy and reliability. This process includes the application of bias correction methodologies to account for any systematic errors in the measurements (Schmidtko et al., 2013).

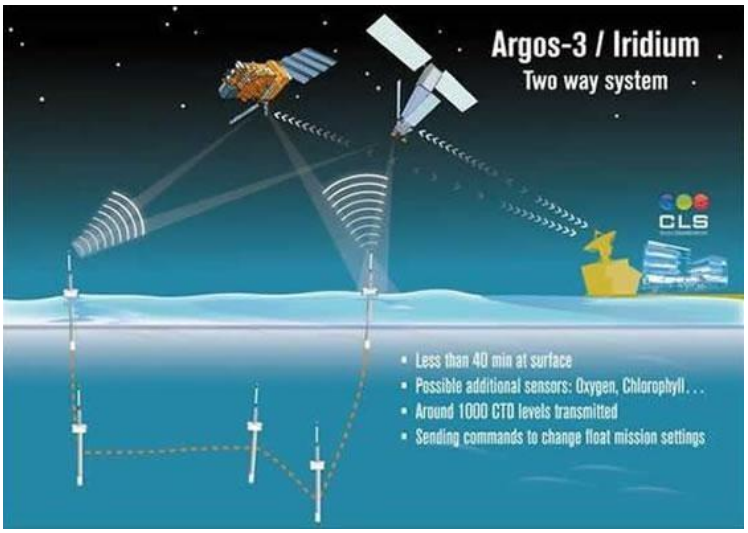


Figure 2 - Data Transmission Process (Argo Program, n.d.)

In addition to temperature and salinity, the ARGO floats also measure pressure, which is essential for calculating the density of sea water and understanding the vertical structure of the ocean. The integration of these measurements allows researchers to analyze ocean stratification, mixing processes, and the dynamics of ocean currents (Mork et al., 2019).

Furthermore, the ARGO project has evolved to include additional capabilities, such as the integration of biogeochemical sensors on some floats. These sensors enable the collection of data on oceanic carbon dioxide levels, oxygen concentrations, and other biogeochemical components, thereby expanding the scope of the ARGO dataset to include information about ocean health and its response to climate change (Stoer et al., 2023) and is used to understand the role of the ocean in the global carbon cycle. It is also used to assess the impacts of climate change on marine ecosystems.

The ARGO program's technological framework also includes advanced data assimilation techniques that combine ARGO observations with satellite data and other in situ measurements. This approach enhances the accuracy of ocean models and allows for better predictions of ocean behavior (Rasp et al., 2018). The use of machine learning and statistical methods to analyze ARGO data has also gained traction, providing new insights into oceanographic processes and improving the classification of oceanic structures (Gagne et al., 2020).

2.4. DATA QUALITY CHALLENGES

2.4.1. ORIGINS OF DATA QUALITY CHALLENGES IN ARGO

As with other industries, the ARGO program faces data quality problems that can affect its scientific useability.

Missing Values

Missing values in the ARGO dataset can arise from:

- Sensor malfunctions;
- Communication failures, or,
- Environmental conditions that prevent data collection.

For instance, if a float experiences a mechanical failure while submerged, it may not transmit data upon surfacing, leading to gaps in the dataset (Claustre et al., 2020). Additionally, extreme weather conditions can hinder the ability of floats to collect and transmit data, resulting in incomplete profiles (G. C. Johnson et al., 2022).

Sensor Drift

Sensor drift is a significant issue in the ARGO dataset, particularly for temperature and salinity sensors. Over time, sensors may become less accurate due to environmental factors or wear and tear, leading to systematic biases in the measurements. For example, the pressure sensors on ARGO floats can exhibit drift, which affects the accuracy of depth measurements and, consequently, the temperature and salinity profiles derived from them (K. S. Johnson et al., 2017). This drift can result in erroneous data that misrepresents ocean conditions.

Noise

Noise in the ARGO dataset can originate from:

- Environmental variability; or,
- Instrument inaccuracies.

Fluctuations in ocean currents or temperature can introduce random errors in the measurements collected by the floats (T. Wang et al., 2018). This noise can obscure true signals in the data, making it challenging to derive meaningful insights from the dataset.

Bias

Bias can occur in the ARGO dataset due to:

- Selection of floats; or,
- Conditions under which data is collected.

If certain regions of the ocean are overrepresented due to a higher density of floats, the resulting dataset may not accurately reflect global ocean conditions (Wong et al., 2020). Additionally, biases can be introduced if the floats are deployed in areas with specific environmental characteristics, for example northern ocean locations near the north pole where the water temperatures are lower, leading to skewed data that does not represent broader oceanic conditions (Barker et al., 2011).

Data Truncation

Data truncation can occur when the ARGO floats do not reach their intended depths due to operational constraints or environmental conditions. Oke et al. (2022) gives the example that if a float is programmed to profile to 2000 meters, but encounters an obstacle or strong currents, it may truncate its data collection prematurely. This can lead to incomplete profiles that do not capture the full vertical structure of the ocean.

Sparse Data

The ARGO dataset can exhibit sparsity in certain regions, particularly in areas with fewer deployed floats. This sparsity can limit the representativeness of the data and consequently reduce the ability to draw comprehensive conclusions about ocean conditions in those regions. Wong et al. (2020) also says sparse data can complicate interpolation efforts, as there may not be enough measurements to accurately estimate values in under-sampled areas.

2.5. TECHNIQUES TO IMPROVE DATA QUALITY IN THE ARGO DATASET

To address these data quality challenges, several techniques have been implemented to enhance the reliability and usability of the ARGO dataset:

Real-Time Quality Control

The implementation of real-time quality control procedures is used to identify and flag erroneous data as it is collected. (Diamanti et al., 2020), for example, describe a suite of quality control tests that automatically flag "bad" profiles from the Biogeochemical-ARGO dataset, allowing for immediate correction or exclusion of unreliable data. This proactive approach helps maintain the integrity of the dataset.

Delayed-Mode Quality Control

Delayed-mode quality control involves a thorough review after data collection. This process corrects sensor drift and biases and also applies statistical techniques to identify errors.

For instance, the delayed-mode salinity data undergoes evaluation. This evaluation adjusts for sensor drift. The adjustments lead to improved accuracy. They also reduce bias in the dataset which results in a more reliable and accurate dataset (Oke et al., 2022). Accurate data is necessary for understanding ocean conditions as it allows scientists to make informed decisions.

The quality control process ensures that the data is usable. It helps in assessing ocean health and climate change and without this process, the data may misrepresent ocean conditions.

In the ARGO dataset, it enhances its overall quality. This process is necessary for effective oceanographic research, because it ensures that the data collected is trustworthy and valid.

Data Assimilation Techniques

Data assimilation techniques, such as the multivariate deterministic ensemble Kalman filter, are usually employed to integrate ARGO data with other observational datasets and model outputs. This approach helps to improve the representation of subsurface properties by combining sparse ARGO profiles with satellite observations and model predictions (B. Wang et al., 2021). By assimilating diverse data sources, researchers can enhance the overall quality and completeness of the dataset.

Enhanced Sensor Technology

The development and deployment of advanced sensor technologies can help mitigate issues related to sensor drift and noise. The integration of hyperspectral technology on ARGO floats has the potential to improve the quality of radiometric observations, enhancing the dataset's overall reliability (B. Wang et al., 2021).

2.6. BREAKTHROUGHS ON DATA QUALITY ENHANCEMENT

In recent years, significant advancements have been made in enhancing the quality of in situ oceanographic measurements, particularly those utilizing Argo floats. These improvements

are used to address data issues such as sensor drift, noise, and missing data, which can compromise the integrity of oceanographic data.

2.6.1. Advanced AI Techniques in Ocean Modeling

Recent advancements in artificial intelligence (AI) have significantly influenced the field of ocean modeling. The application of machine learning techniques, particularly deep learning, has shown promise in enhancing the accuracy of oceanographic predictions.

For instance, Fu (2014) conducted a genetic diversity analysis of highly incomplete genotype data, demonstrating the effectiveness of machine learning methods in managing missing data. The study revealed that employing sophisticated imputation techniques could substantially improve the accuracy of genetic diversity assessments. This finding is relevant to oceanographic research, where machine learning approaches can be utilized to analyze complex datasets and impute missing values, thereby bolstering the reliability of ocean models.

Moreover, Ernst & Kellis (2015) discussed the large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. The authors highlighted the challenges associated with imputing transcription factor binding data, which can vary significantly across samples. This variability mirrors the challenges faced in oceanographic datasets, where environmental factors can influence the measurements collected. The study emphasizes the necessity for robust imputation frameworks that can account for the inherent variability in data, ultimately enhancing the quality of analyses and predictions.

2.6.2. Evolution of Big Data Analytics and Machine Learning Approaches

The evolution of big data analytics has transformed the methodologies employed by researchers analyzing large-scale oceanographic datasets. Qian et al. (2022) introduced Informer-WGAN, a high missing rate time series imputation method based on adversarial training and a self-attention mechanism. Their findings demonstrated that this approach could effectively manage missing data in time series, providing a valuable tool for researchers dealing with incomplete datasets. The ability to impute missing values while preserving the temporal structure of the data is crucial for understanding ocean dynamics and climate variability.

In a related study, Folguera et al. (2015) examined the use of SOM for the imputation of missing data in incomplete data matrices. The authors highlighted the advantages of SOMs in modeling nonlinear relationships within environmental datasets, making them suitable for oceanographic applications. By leveraging the relationships among pertinent variables, SOMs can provide accurate estimates for missing values, thereby enhancing the overall quality of datasets. This study underscores the potential of machine learning techniques to improve data quality and facilitate more accurate analyses in oceanography.

2.6.3. Machine Learning Algorithms

Machine learning techniques have proven to be effective for ARGO data. They help researchers maintain the integrity of the dataset by addressing data quality issues, such as missing values.

These methods enhance usability, leading to better insights into ocean dynamics and climate change. Fazakis et al. (2020) states the integration of machine learning in oceanographic research has improved the understanding of complex ocean processes.

One commonly used method is K-Nearest Neighbors (KNN). KNN imputes missing values by finding the nearest data points and using these neighbors to predict missing values and, as an example, has been successfully applied in solar energy forecasting to handle missing data (Kim et al., 2019).

Another popular technique is Random Forest (RF). This method is robust and handles complex relationships, making it particularly useful for high-dimensional data. In a study on missing data in household travel surveys, RF outperformed traditional methods (Budhwani et al., 2023), proving is reliability when compared to traditional methods.

Support Vector Machines (SVM) have also been applied for imputation. This method is beneficial for datasets with non-linear relationships, such as the ARGO dataset and has been used in software development effort estimation to address missing data (Idri et al., 2018).

Additionally, Multivariate Imputation by Chained Equations (MICE) is another method. MICE creates multiple datasets with different imputed values and then combines these datasets to provide a final estimate. This approach is useful for handling uncertainty in missing data and has been widely used in ecological studies to improve the accuracy of trait databases (Penone et al., 2014).

The MissForest algorithm combines Random Forest with an iterative approach. It imputes missing values by using Random Forest predictions and has shown superior performance in various studies. This algorithm outperformed others, such as KNN and MICE, in imputing missing data in life-history trait datasets (Penone et al., 2014). It is particularly effective for mixed-type data, which is common in ARGO datasets.

2.6.4. Challenges in Handling Large-Scale Datasets

Despite the advancements in spatial-temporal data mining and machine learning applications, several challenges remain in managing large-scale oceanographic datasets. (Jo, 2022)

emphasized the importance of retaining and quantifying uncertainty in imputation procedures, particularly in the context of multibeam angular response data for seabed mapping. This focus on uncertainty is critical in oceanographic research, where data quality can significantly impact the reliability of analyses and predictions.

Future research should concentrate on developing more sophisticated algorithms capable of effectively handling the complexities of ocean data, including non-linear relationships and high dimensionality. Additionally, integrating multi-source data, such as satellite and in-situ measurements, can enhance the robustness of predictive models and improve the accuracy of ocean forecasts. Ongoing efforts to enhance data quality through advanced preprocessing techniques and imputation methods will be essential for ensuring the reliability of analyses and predictions in oceanographic research.

2.7. CONCLUSION

This review of the literature examines the historical and technical parts of the ARGO program as well as the common problems associated with data quality and how they are likely to be generated.

In addition, it explores techniques to address different ways of dealing with these challenges, as well as machine learning algorithms and other ways of solving data quality problems.

The following chapter presents methods for examining and correcting data quality problems in ARGO dataset.

3. METHODOLOGY

3.1. INTRODUCTION

This study adopted a quantitative approach to analyze and impute missing values in the ARGO float dataset, a global oceanographic repository of in situ measurements. The methodological framework was grounded in the CRISP-DM (Cross-Industry Standard Process for Data Mining) paradigm, which was adapted to fit the specific needs of this research.

The process was divided into seven structured phases, as illustrated in Figure 3: Data Collection, Data Understanding, Data Transformation, Data Exploration, Data Preparation, Model Training and Testing, and Evaluation.

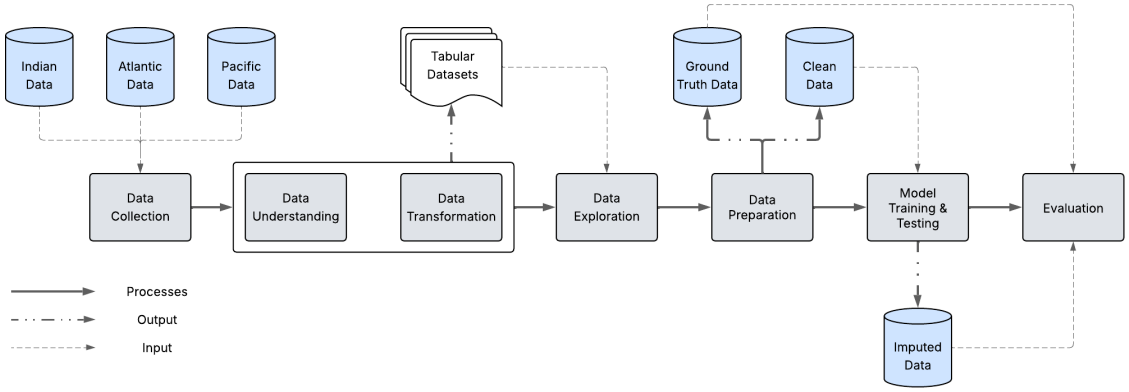


Figure 3 – Methodology Workflow

3.1.1. Workflow WalkThrough

Data collection was carried out using a loop that extracted files from the ARGO repository and saved them locally. The data was provided in NetCDF format, which is specifically designed for storing and sharing multidimensional scientific datasets, such as oceanographic observations, in a compact, self-describing, and platform-independent manner.

The next step was to understand and transform the data. These two processes were made simultaneously and consisted of transforming the data from NetCDF format to a tabular database that could be properly analyzed and understood by the models.

Following is Data Exploration, where the focus shifted toward understanding the structure, quality, and underlying patterns of the dataset. To support this process, several visualizations were created to help uncover trends, identify anomalies, and gain deeper insights into spatial and oceanographic variables.

Next, the data was prepared for model training through a series of preprocessing steps. This included removing null values, filtering out outliers, and applying normalization to ensure consistency across feature scales, along with a few other minor adjustments. Once a clean

dataset was established, a controlled level of missingness was artificially introduced to simulate real-world scenarios. The original (ground truth) values were stored separately to allow for accurate comparison during the evaluation phase.

3.2. DATA COLLECTION

Extracting, processing, and storing data required a structured approach to maintain consistency and reliability. This dataset included only physical parameters collected globally by Argo profiling floats.

To maintain consistency, the data extraction period is predefined from January 1st, 2019, to December 31st, 2019. Proper date formatting ensured that queries to the ARGO repository retrieved accurate and aligned records.

Given the vast scale and structured layout of the archive, a custom web scraping tool was developed in Python to automate the ingestion process. The script looped over the predefined list of oceans, these being the Pacific, Atlantic and Indian, and year of interest. For each combination, it constructed the correct web path to the corresponding data folder on the GDAC FTP server. Each monthly subfolder was then parsed to extract links to all valid NetCDF files matching the *nodc_R* naming pattern, which identifies raw profile data.

To process the directory listings, the script used a library to retrieve the HTML content and *BeautifulSoup* parsed and filtered relevant file links. Once the file list was obtained, it proceeded to download them sequentially.

A key feature of the implementation was its robustness as before attempting to download any file, the script checked whether it already exists locally, thus preventing redundant downloads and saving bandwidth. Additionally, robust exception handling ensured that errors in individual folders (e.g., broken links or missing directories) were logged, but did not interrupt the overall process.

Once extracted and verified, the data was stored in local folders. A structured local storage ensured easy accessibility and compatibility with analytical tools. Filenames included timestamps or float IDs for proper organization, for example the file regarding March 2019 was saved as *ARGOData_202403.csv*. Encoding and delimiter settings were standardized to prevent format-related errors, ensuring data integrity across different processing environments.

This automated data ingestion approach resulted in 84930 files, or 84930 cycles as each file corresponds to one cycle, with the structure presented on Table 2.

Variable Name	Description	Data Type
ocean	Ocean basin where the profile was collected	string
year	Year when the profile was collected	integer
juld	Timestamp of the profile measurement	datetime
latitude	Latitude of the observation (in decimal degrees)	float
longitude	Longitude of the observation (in decimal degrees)	float
cycle	ARGO float cycle number	integer
pressure	Measured pressure at that depth (in decibars)	float
temperature	Measured temperature (in degrees Celsius)	float
salinity	Measured salinity (in PSU - Practical Salinity Units)	float

Table 2 - Raw NetCDF Files Variable Information

3.3. DATA UNDERSTANDING

These files are self-describing and platform-independent, making them particularly well-suited for handling structured oceanographic datasets. Each file was organized into dimensions, variables, and attributes (Table 3).

File Features	Description
Dimensions	Define the axes over which data is recorded (e.g., time, depth, latitude, longitude).
Variables	Store the actual measurement values. It can be associated with one or more dimensions.
Attributes	Metadata attached either to variables or to the dataset as a whole.

Table 3 - Structural Components of NetCDF Files

This structure allowed for efficient storage of large volumes of profile-based observations, while preserving the contextual information necessary for analysis. In the context of the ARGO program, each NetCDF file often contained one or two vertical profiles, with measurements taken at varying depths and referenced by time and geographic coordinates.

For this study, only the second profile in each ARGO file was analyzed, as it corresponds to the ascending profile, the phase during which the float returns to the surface and typically collects the most reliable measurements. In cases where a file contained only one profile, it was treated as the second (ascending) profile by default.

3.4. DATA TRANSFORMATION

The transformation process began by selecting a sample NetCDF file to explore its structure, metadata, and variable dimensions. This exploratory step was useful to identify the naming conventions used across ARGO profiles and the data itself in a tabular format.

The script looped through each profile and depth level, extracting relevant metadata and geophysical values. Invalid or masked entries were filtered out using built-in masking mechanisms provided by the *NetCDF4* and *NumPy* libraries. Where pressure was available, depth, in meters, was calculated using the pressure values, adding an additional dimension to the analysis.

To ensure traceability, each row in the resulting dataframe retained the original source filename, ocean basin, and year of acquisition. This metadata was used later for aggregation and comparative oceanographic analysis.

Two distinct output files were generated per ocean: one containing only complete profiles (**_Profiles_YYYY.csv*) and another containing only masked entries (**_Masked_Only_YYYY.csv*). These were merged into a single file per ocean, and finally, a global dataset (Table 4) was produced containing approximately 45 million rows and 11 columns regarding the observations from the Atlantic, Pacific, and Indian Oceans.

source_file	ocean	year	juld	lat	long	depth	cycle	pres	temp	sal
nodc_R1901155_317.nc	Indian	2019	2019-09-18T04:49:46.999999	49.822	42.180999	4.3623	317	4.4	2.6	33.946
nodc_R1901155_317.nc	Indian	2019	2019-09-18T04:49:46.999999	49.822	42.180999	8.3107	317	10.4	2.6	33.946

Table 4 - - Final Dataset Preview

This structured and reproducible transformation step ensured that the raw data was successfully flattened and integrated into a consistent schema.

3.5. DATA EXPLORATION

Before diving into model development, a comprehensive exploratory data analysis (EDA) was conducted to understand the structure, quality, and patterns in the oceanographic dataset.

A total of 44.808.101 records were initially compiled. However, the dataset presented significant proportions of missing values, as it is possible to verify on Table 5. The fields that are not represented in the table below had no missing values. Additionally, the dataset was mostly numeric, with key variables represented as floating-point values.

Variable	Missing Count	Missing Percentage
Salinity	25458540	56.82
Pressure	23050334	55.91
Temperature	25034264	55.97

Table 5 – Missing Values per Variable

From a statistical perspective, temperature values ranged from -2.33°C to 39.69°C, with a mean of approximately 6.87°C. Salinity values were generally stable, centered around 34.6 PSU, while depth values exhibited a wide spread, from surface-level observations to measurements taken at over 6,000 meters below sea level.

Figure 4 shows a drill-down version of the overall records, segregated by ocean.

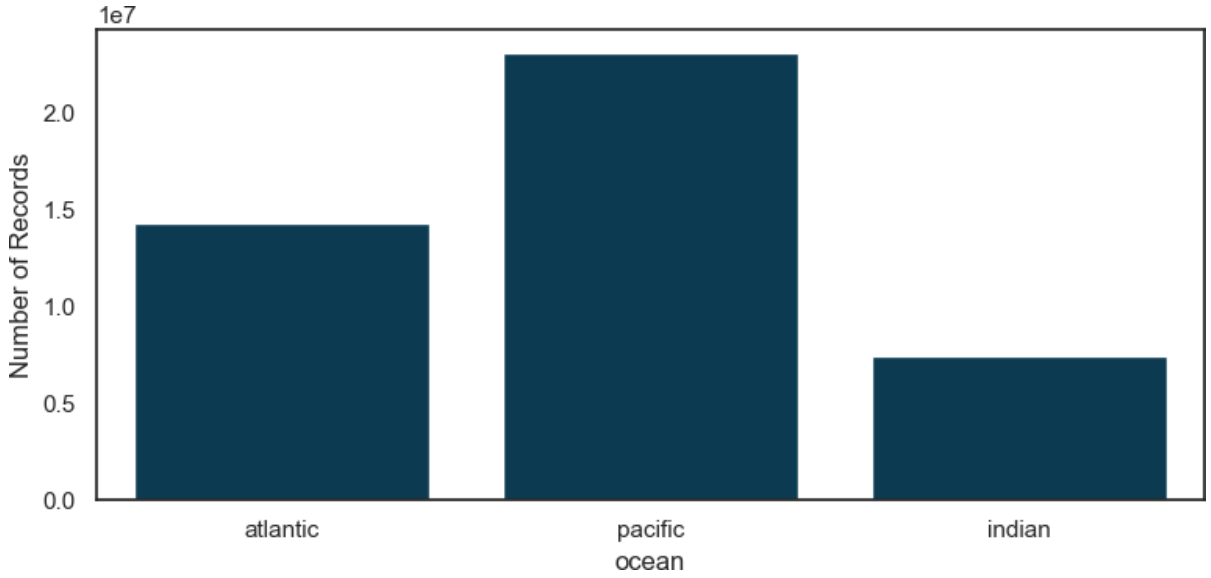


Figure 4 - Number of Records per Ocean

The next step was examining the number of observations by month. As shown in figure 5, there is a clear increase in the number of records throughout 2019.

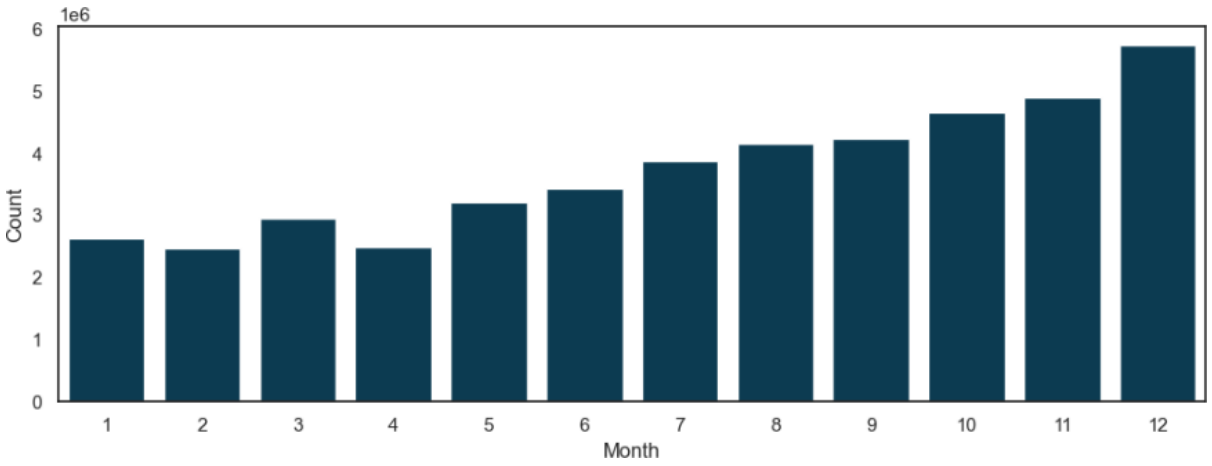


Figure 5 – Number of Records per Month

While the initial months (January to May) registered between 2.5 and 3.5 million observations, a consistent upward trend is evident, peaking at nearly 6 million in December. This may reflect operational variations in ARGO float deployments or seasonal influences on data availability. Understanding this imbalance is important, as it may introduce bias in temporal analyses or models sensitive to time trends.

Building on this, the spatial distribution of data points is illustrated in Figure 6.

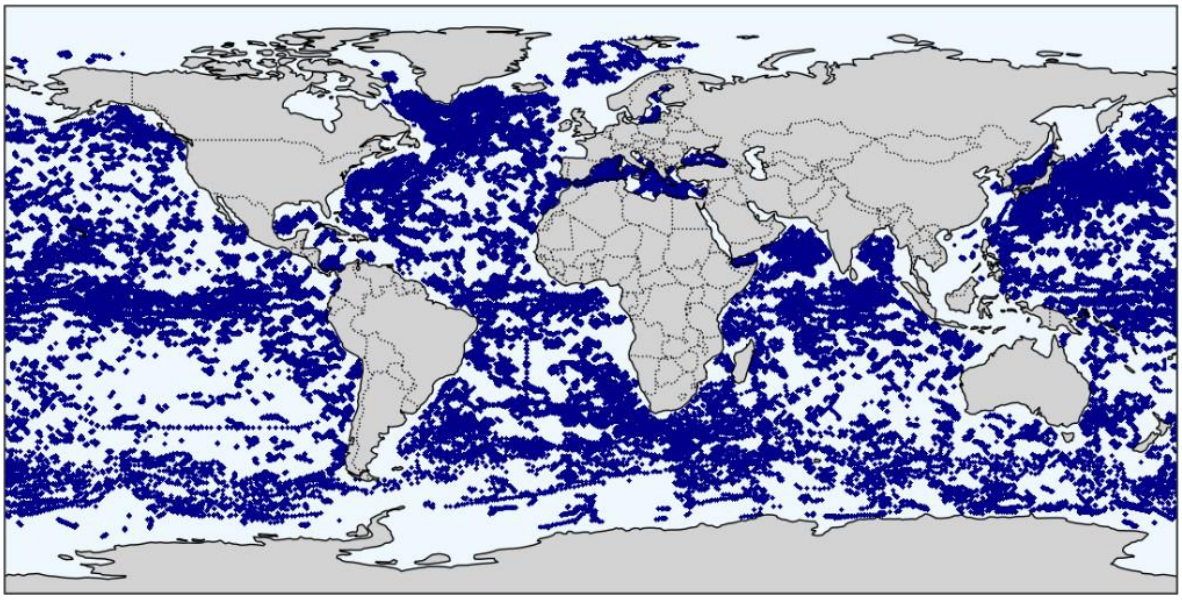


Figure 6 - Float Cycles Spatial Distribution

Here, a global map highlights the extensive geographic coverage of the ARGO dataset, with dense concentrations of observations, i.e. cycles, across all three major oceans: the Atlantic, Pacific, and Indian. Nonetheless, the polar regions and coastal zones remain sparsely represented, pointing to known limitations of ARGO float technology in ice-covered or shallow waters.

To complement the horizontal view, Figure 7 shifts focus on the vertical dimension, showing how observation frequency changes with depth.

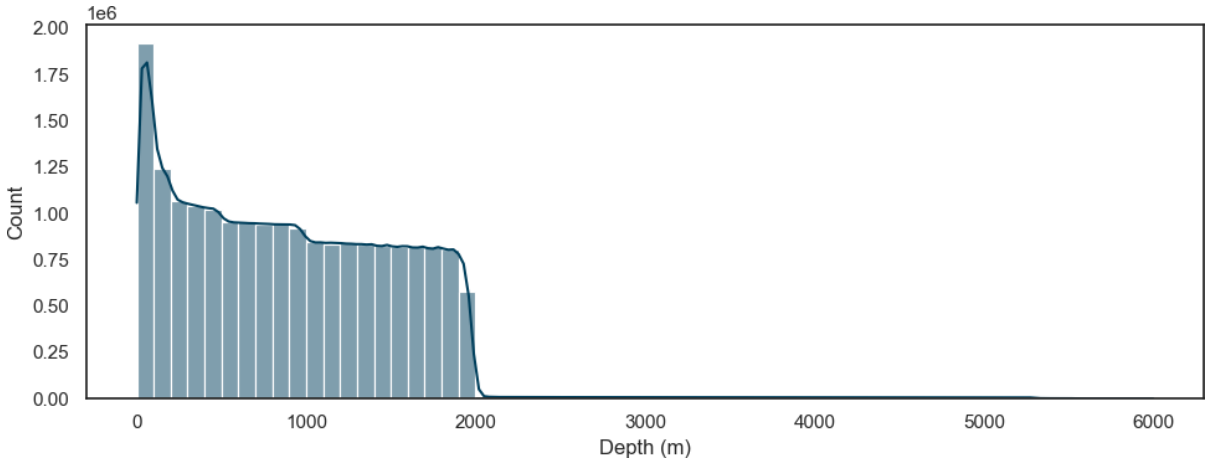


Figure 7 – Records per Depth Level

The vast majority of measurements are located within the upper 2000 meters, with a sharp drop-off beyond that point. This aligns with the ARGO program's mission to monitor upper-ocean properties, while also highlighting the depth-related sparsity that must be considered in modeling efforts.

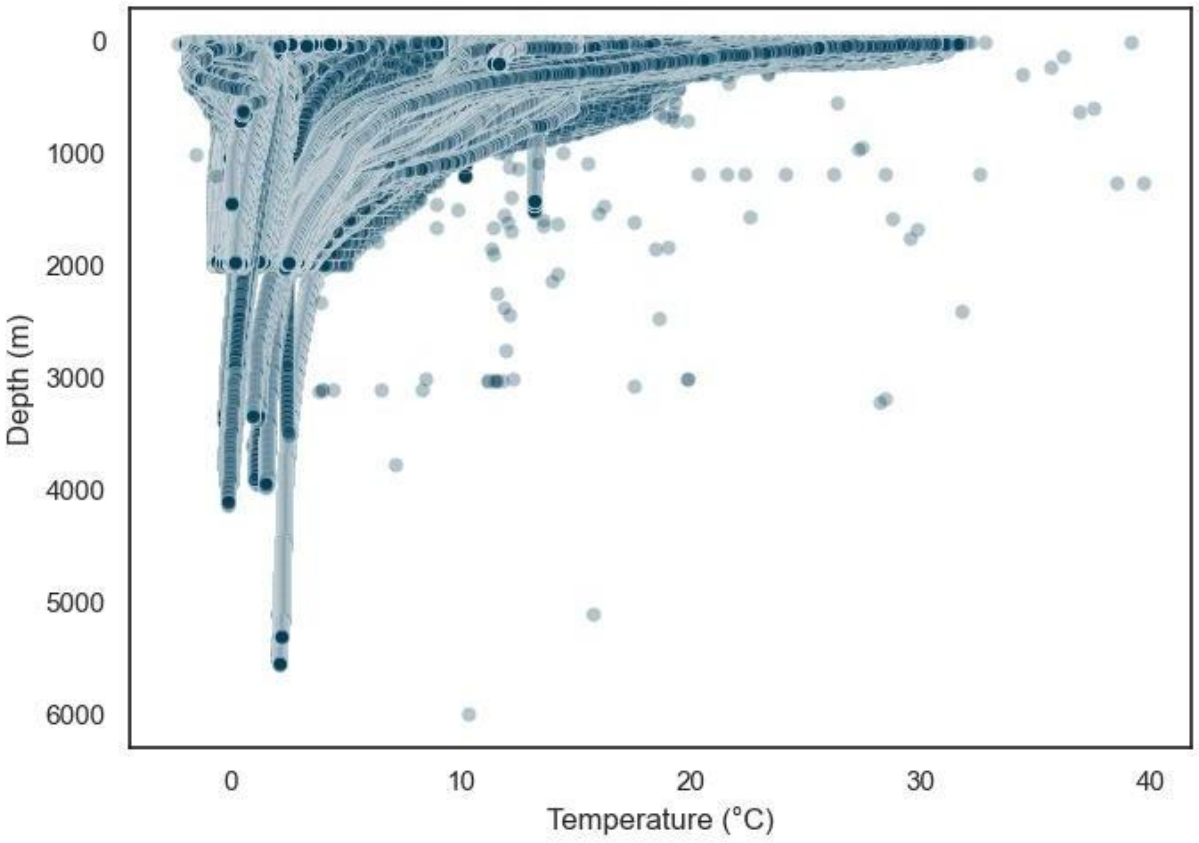


Figure 8 – Temperature vs Depth Relationship

Further insight into the structure of the dataset comes from the temperature–depth relationship. As expected, a clear stratification pattern emerges as high temperatures are concentrated near the surface, decreasing with depth as it is possible to verify below.

To deepen this understanding, pairwise relationships (Figure 9a) between variables are explored. While temperature shows a strong negative correlation with both pressure and depth, salinity displays a weaker and more scattered relationship. These patterns are quantified in the correlation heatmap (Figure 9b), where the temperature–depth and temperature–pressure relationships are confirmed with correlation coefficients around -0.65. In later phases of the methodology, these relationships and correlations will be useful to reduce the dimensionality of the dataset. Please note that the correlation between pressure and depth is 1.00 as depth was calculated based on the pressure column.

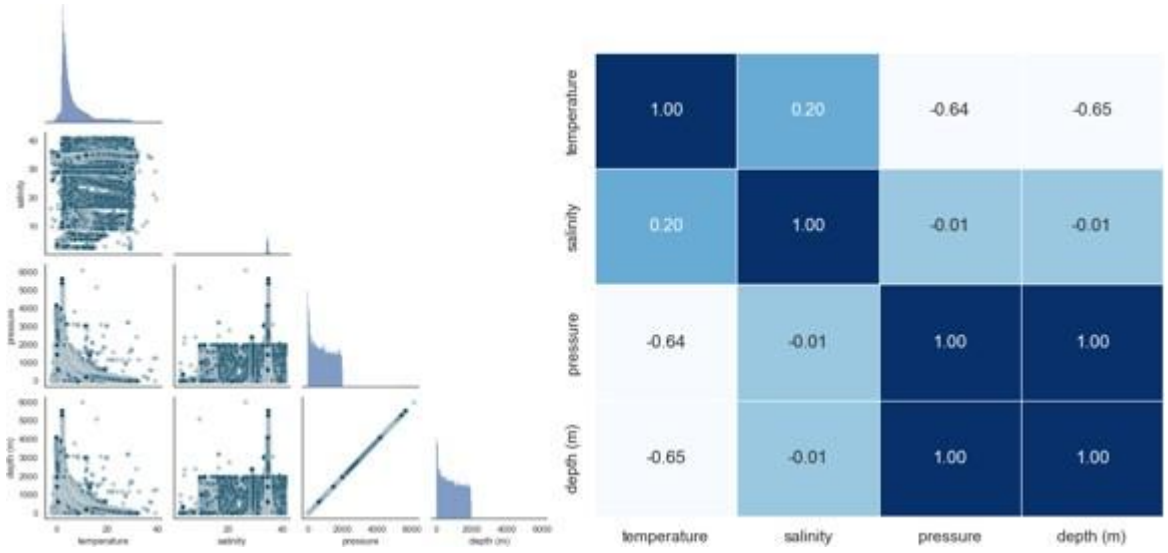


Figure 9 - a) Overall Variables Shape b) Pearson Correlation

Finally, an essential step in this analysis involved identifying outliers and unusual patterns in the data, as they need to be removed in the modelling phase. Figure 10 provides boxplots for temperature, salinity, and pressure.

These analyses were later used as insights in the modelling phase and some key decisions were taken based on the EDA.

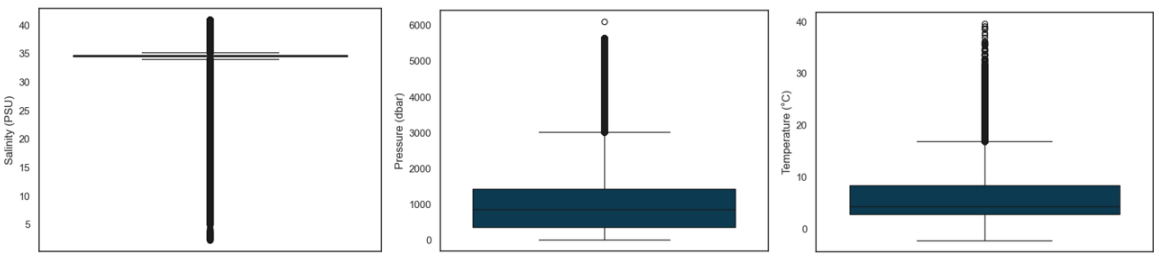


Figure 10 - Key Variable Patterns

3.6. DATA PREPARATION & TRAINING

To improve data quality, a series of preprocessing steps were applied. Outlier detection and treatment were particularly important to address potential issues such as sensor malfunctions, extreme environmental events, or data transmission errors. These anomalies can introduce bias, distort statistical distributions, and negatively impact model performance. One of the main techniques used was Z-score analysis, which identifies extreme deviations from the mean by flagging observations with values exceeding a defined threshold (typically ±3 standard deviations). This method ensures that the dataset remains representative of realistic oceanographic conditions while reducing the influence of outlier-driven noise.

To prepare the data for training the ML models, several transformations were applied to ensure consistent scaling and prevent feature dominance.

First, the pressure variable, having a significantly larger range than the other features, was log-transformed, which compressed its scale while preserving relative differences. This transformation helped reduce skewness and brought the distribution closer in line with other variables. Afterward, all input features were normalized using *MinMax* scaling, which mapped values to a range between 0 and 1. This normalization was applied to both the spatial features (latitude and longitude) and the oceanographic variables (temperature, salinity, and log-transformed pressure). These steps were applied, as the models rely on distance-based learning, ensuring that all features contribute equally and that no single variable dominates due to its magnitude.

Additionally, when working with multiple oceanographic parameters, some variables may be redundant or highly correlated, which can introduce unnecessary complexity or multicollinearity into the model. For this reason, feature selection was applied to reduce redundancy and improve model interpretability. In particular, the depth variable was removed because it was found to be perfectly correlated with pressure, meaning it did not contribute any additional information and could be safely excluded without loss of predictive value.

Finally, to ensure data accuracy, three separate subsets were created:

- **Training Dataset**: This dataset consists of 80% of all data where the original null values were removed.
- **Testing Dataset:** This dataset consists of the remaining 20%. Additionally, 50% of these values were randomly removed to replicate original missingness.
- **Groud-truth Dataset**: This dataset stored the original non-null values that were removed at random from the Testing Dataset.

This dataset segregation approach was used to address inconsistencies in the data caused by sensor errors, data transmission failures, or historical inaccuracies and allowed to truly verify the ML models accuracy in the evaluation phase.

3.7. MODELS EXPLANATION

The ARGO dataset contains oceanographic measurements with gaps caused by sensor failures, environmental conditions, and transmission errors. To address these gaps, the data imputation methods presented on Table 6 were applied.

Method	Approach	Application in ARGO Data	Role
SOM	Unsupervised pattern clustering	Imputes missing ocean properties	Main Model
KNN	Distance-based estimation	Estimates missing values using neighboring profiles	Baseline Model

Table 6 – Data Imputation Methods

3.7.1. Self-Organizing Maps

Self-Organizing Maps are a class of unsupervised artificial neural networks designed to project high-dimensional input data onto a low-dimensional (typically 2D) grid while preserving topological and neighborhood relationships. This property makes SOMs particularly suited for exploratory data analysis, clustering, and, as applied in this thesis, imputation of structured environmental data such as oceanographic profiles.

At its core, the SOM model consists of a discrete lattice of neurons, where each neuron i is associated with a weight vector $w_i \in \mathbb{R}^n$, with n representing the dimensionality of the input space. These weight vectors evolve during training to represent the manifold structure of the data. The network learns to organize itself by iteratively adjusting the neurons' weights in response to incoming input vectors x, without requiring labeled targets or explicit supervision.

Grid Topology

The SOM grid can be rectangular or hexagonal (Figure 11), and the dimensionality is usually 2D to facilitate visualization and interpretation. In this exercise, the rectangular grid was used.

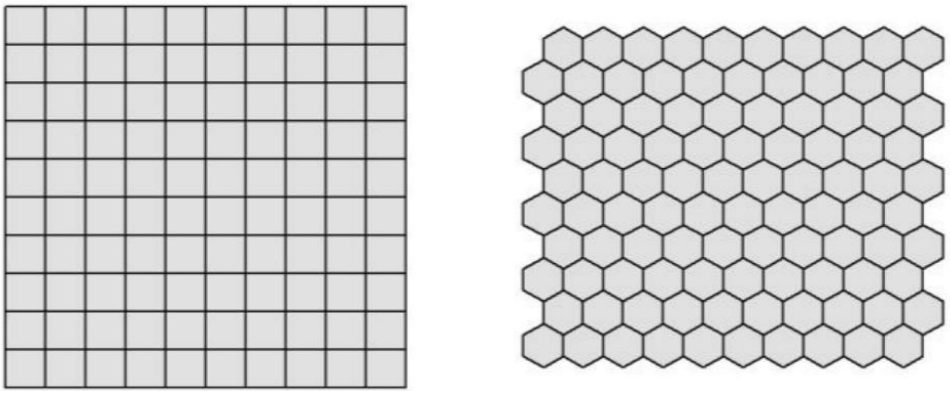


Figure 11 – a) SOM Rectangular Grid b) SOM Hexagonal Grid

The grid's geometry defines the neighborhood relationships between neurons and plays a central role in how local and global patterns are captured. Each neuron has a fixed location in the grid, and its influence on surrounding neurons is controlled by a neighborhood function.

Training Procedure

The learning algorithm proceeds iteratively, typically through the following steps:

1. **Initialization**

   Weight vectors $w_i(0)$ are initialized. This can be done randomly or by sampling from the input space.

2. **Competition: Find the Best Matching Unit (BMU)**

   For each input sample x, the network identifies the neuron whose weight vector is closest in terms of Euclidean distance:

$$BMU = \arg\min_i ||x - w_i|| \tag{1}$$

   This neuron represents the cluster that is most similar to the input vector.

3. **Adaptation: Updating Weights**

   The BMU and its neighboring neurons are updated to become more similar to the input vector. The weight update rule at time t is given by:

$$w_i(t+1) = w_i(t) + \eta(t) \cdot h_{b,i}(t) \cdot (x - w_i(t)) \tag{2}$$

   where:

- $\eta(t)$ is the learning rate, which decays over time.
- $h_{b,i}(t)$ is the neighborhood function centered around the BMU b, often modeled as a Gaussian:

$$h_{b,i}(t) = \exp\left(\frac{\| r_b - r_i \|^2}{2\sigma(t)^2}\right) \tag{3}$$

   Here, r_i and r_b are the positions of neurons i and the BMU in the grid, and $\sigma(t)$ controls the spread of the neighborhood and also decays with time.

4. **Convergence**

   This process is repeated over many epochs, typically until the weight vectors converge or until the maximum number of iterations is reached.

Missing Data Handling and Imputation

After training, SOM represents a structured manifold of the input space. Each input vector with missing values can be projected onto the SOM by identifying its BMU using only the non-missing features. The corresponding complete vector can then be reconstructed based on:

- The BMU's weight vector.
- A weighted average of neighboring neurons (optional).
- Multiple BMU candidates, in case of ambiguity or partial matches.

This process allows for the estimation of missing variables by leveraging the patterns learned during unsupervised training, effectively using the structure of the data to infer unknown values. The preservation of topological relationships ensures that imputations remain consistent with regional and environmental patterns in the original dataset, especially relevant for oceanographic features with spatial and vertical dependencies.

3.7.2. K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a non-parametric, instance-based learning algorithm primarily used for classification and regression. However, it is also a widely adopted technique for data imputation, particularly when the dataset contains partially missing entries in a tabular format. The method leverages the intuition that similar samples tend to have similar values, and it infers missing values by examining the most similar observations in the dataset.

In the context of imputation, KNN operates by identifying the k most similar samples that have no missing values for the target features. The similarity between instances is usually computed using a distance metric, such as Euclidean distance, across the observed features. Once neighbors are identified, the missing value is estimated by aggregating the values of these neighbors using a specific strategy—such as a mean, median, or distance-weighted average.

Training Procedure

Formally, given an incomplete instance $x \in \mathbb{R}^n$ with some features missing, the KNN imputation method involves:

1. **Distance Calculation**

   Compute the distance between x and all other complete instances using the non-missing features.

2. **Neighbor Selection**

   Identify the k instances $\{x_1, x_2, \ldots, x_n\}$ with the smallest distances.

3. **Value Estimation**

For each missing feature, compute a weighted average of the corresponding values from the neighbors:

$$\hat{x}_{missing} = \frac{\sum_{i=1}^{k} wi \cdot x_{i,missing}}{\sum_{i=1}^{k} wi} \tag{4}$$

where the weight wi is typically defined as the inverse of the distance between x and x_i:

$$w_i = \frac{1}{distance(x,\ x_i) + \varepsilon'} \tag{5}$$

with ε being a small constant to prevent division by zero.

3.8. DATA IMPUTATION

3.8.1. Self-Organizing Maps

After the SOM was trained, missing values were reconstructed by leveraging the map's topological structure. Each incomplete observation was first assigned to its BMU, and imputation was then performed using the BMU's weight vector or a weighted aggregation of its neighbors' vectors. The imputed value x' is computed as:

$$x' = \frac{\sum_{i \in N(BMU)} h(i) w_i}{\sum_{i \in N(BMU)} h(i)} \tag{6}$$

where:

- w_i represents the weight vector of neuron i.
- $h(i)$ is the Gaussian neighborhood kernel.
- $N(BMU)$ denotes the set of neurons within a fixed radius of the BMU.

The SOM was trained in the complete feature space using stochastic learning over 100,000 iterations. The training employed a Gaussian neighborhood function, with a learning rate of 0.3 and a sigma of 1.85, selected to balance sensitivity to local gradients with global topological coherence. The grid dimensions were determined heuristically from the dataset size and dimensionality to ensure sufficient granularity for meaningful clustering.

Following training, each complete record was matched to a BMU, and observations were grouped accordingly to construct a mapping dictionary that enabled efficient lookup of contextually similar data points.

For imputation, incomplete records were initially processed by temporarily substituting missing entries with their respective global means. This ensured a consistent feature vector

for BMU assignment. After identifying the BMU, the algorithm retrieved all neurons within a 1-unit radius and extracted the associated observations from the mapping dictionary.

If no valid neighbors were found—e.g., in sparsely populated regions—the algorithm defaulted to the global mean for the missing variable. Otherwise, the imputation relied on inverse-distance weighting, computed from geographic coordinates (latitude and longitude), giving more influence on nearby neighbors. This design choice reflects the spatial continuity inherent in oceanographic variables, where local conditions often dominate over distant ones.

The final imputed value for each missing element was derived as a weighted average over the identified neighbors. This operation was applied row-by-row to the artificially masked test dataset, and the imputed results were retained in a separate copy for subsequent performance evaluation.

3.8.2. K-Nearest Neighbors

The KNN imputation approach estimated missing values based on the similarity between observations in the dataset. It assumed that instances with similar known features were likely to share similar values for their missing features. This made KNN particularly suitable for datasets like ARGO, where measurements such as temperature, salinity, and pressure vary smoothly across space and time.

The process involved three main steps: training the imputer, applying it to the masked dataset, and transforming the results back to the original scale.

A *KNNImputer* was first fitted to a version of the dataset where all variables were complete and preprocessed. In this step, the imputer calculated pairwise distances between all observations in the training set. These distances were computed in a normalized feature space to ensure comparability across variables with different units and ranges. In this case, standard feature scaling was applied using a *MinMaxScaler* prior to fitting, so that each feature had zero mean and unit variance.

The imputer stored the structure of the dataset in this scaled space, which enabled efficient lookup of the k = 5 nearest neighbors for any new observation with missing values.

The dataset with artificially masked entries (representing missing values) was then passed to the trained imputer. For each incomplete record, the algorithm identified the 5 nearest neighbors (other observations that were complete) using Euclidean distance in the same scaled feature space. Only features that were present in both the target and the neighbor observations were used in the distance calculation, ensuring that missing entries did not bias similarity computation.

For each missing feature, the imputer computed a distance-weighted average of the corresponding feature values from the selected neighbors. Closer neighbors had a higher influence on the imputed value than distant ones. This approach enhanced local consistency, as it prioritized the behavior of nearby or similar oceanographic profiles in estimating missing measurements.

After imputation, the completed dataset still resided in the scaled feature space. To recover the original measurement units, the imputed data was passed through the inverse transformation of the original *scaler*. This step ensured that the imputed values were physically interpretable and aligned with the rest of the ARGO dataset.

Finally, the reconstructed dataset was stored in a new data frame for subsequent evaluation and comparison with other imputation methods.

3.9. MODEL EVALUATION

The final stage of the methodology involved evaluating the accuracy of the imputation models, in order to determine whether the reconstructed values could reliably replace missing data in operational or scientific workflows.

The predicted values were compared to the original data using the following evaluation metrics:

Mean Absolute Error (*MAE*)

MAE quantified the average absolute difference between the true and imputed values. It is defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\, y_i - y_i|$$

(7)

where:

- n is the total number of imputed values,
- y_i is the true value for observation i,
- $\hat{y}_i$ is the corresponding imputed value.

MAE is a linear metric, treating all errors equally regardless of their magnitude.

Root Mean Squared Error (*RMSE*)

RMSE provided a measure of the average magnitude of the error, but gives more weight to large errors due to the squaring term:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

(8)

This metric is particularly useful when large deviations are more detrimental than small ones, which is often the case in physical datasets like ARGO.

Coefficient of Determination (R^2)

R^2 measured how well the imputed values approximate the true data distribution. It is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

(9)

where:

- $\bar{y}$ is the mean of the true values y_i.

R^2 ranges from negative infinity to 1. A value of 1 indicates perfect imputation, while values closer to 0 suggest weak predictive power. Negative values indicate that the model performs worse than simply using the mean of the observed values.

*Mean Absolute Percentage Error (*MAPE*)*

MAPE evaluates the accuracy of regression models, especially in forecasting or imputation tasks. It expresses the average error as a percentage of the actual values.

Given a set of true values $\{y_1, y_2, \ldots, y_n\}$ and their corresponding predicted or imputed values $\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}$, the *MAPE* is defined as:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

(10)

where:

- y_i is the true value,
- $\hat{y}_i$ is the imputed value,
- n is the total number of observations.

3.9.1. Evaluation Scopes

Variable-Specific Evaluation

Each metric was computed separately for temperature, salinity, and pressure to account for the differences in variability and scale. For instance, pressure was log-transformed during preprocessing, which may affect interpretation. This stratified approach ensured a more nuanced understanding of the model's behavior across different physical dimensions of the ocean.

Together, these four metrics offer a complementary view of model performance. *MAE* and *RMSE* describe the magnitude of the errors, while R^2 captures how well the structure of the original data is retained. *MAPE* provided an intuitive percentage-based measure of error relative to the true values, enabling scale-independent comparisons across variables.

Ocean-Specific Evaluation

In addition to evaluating performance by variable, the imputation model was assessed across distinct oceanic regions to account for spatial heterogeneity in oceanographic processes. The dataset was geographically stratified into major ocean basins—such as the Atlantic, Pacific, and Indian Oceans—as well as polar and equatorial zones where data density and physical dynamics differ markedly.

For each region, the same evaluation metrics were computed independently. This allowed for the identification of spatial biases or weaknesses in the imputation framework. For instance, under-sampled regions like the Southern Ocean often exhibit greater error due to sparser observational coverage and more extreme environmental conditions. Conversely, more densely observed regions tended to yield more stable and accurate imputations.

Segmenting the evaluation by region also provided insight into the effectiveness of the SOM's spatial generalization capabilities. Because the model leveraged both latitude and longitude during training, its ability to interpolate missing values while preserving geographical structure was a critical aspect of its performance. By quantifying errors within each ocean segment, the evaluation helped clarify where the model retained topological coherence and where its assumptions—such as spatial continuity—might have broken down.

This region-specific analysis therefore complemented the variable-wise assessment, offering a multidimensional view of the model's strengths and limitations across both physical and geographic domains of the ocean.

4. RESULTS & DISCUSSION

This chapter presents the results and discussion for each variable at a global level and at an ocean level.

4.1. GLOBAL PERFORMANCE EVALUATION

The evaluation was carried out using four common metrics for regression problems, as mentioned in the methodology: Mean Absolute Error, Root Mean Squared Error, the coefficient of determination and the Mean Absolute Percentage Error. These metrics allowed for a nuanced assessment of imputation accuracy, ranging from overall deviation (*MAE*), penalization of large errors (*RMSE*), and explanatory power of the model (R^2), to relative error expressed as a percentage of the true value (*MAPE*).

Feature	MAE	RMSE	R^2	MAPE (%)
Temperature	1.7944	3.8202	0.2904	29.0083
Salinity	0.0962	0.3116	0.5702	0.2741
Pressure*	0.4124	0.7324	0.3697	7.2970

Table 7 - SOM's Results by Target Variable

The Self-Organizing Map model demonstrated strong performance in imputing salinity data, achieving a *MAE* of 0.0962 PSU, *RMSE* of 0.3116, and a low *MAPE* value of 0.2741%. These values indicate high precision and consistency, which can be largely attributed to the inherently low variability of salinity in open ocean environments. As confirmed by the boxplot distribution, most salinity values fall within a narrow range (~34–36 PSU), offering a stable structure for the SOM to learn from. Additionally, salinity varies more gradually with depth and across latitudes than other variables, allowing the SOM to effectively cluster and interpolate across similar profiles. The combination of depth, latitude, and longitude in the SOM's input space further enhanced this coherence, enabling the projection of data into a low-dimensional grid that preserves the topological relationships between water masses.

In contrast, temperature posed a greater challenge for the model. The SOM model yielded a significantly higher *MAE* of 1.7944°C, *RMSE* of 3.8202°C, and a low R^2 of 0.2904, with *MAPE* exceeding 29%. These weaker results reflect the high spatial and vertical heterogeneity of ocean temperature, which is subject to short-term variability, seasonal cycles, thermocline structures, and mixed layer dynamics. The vertical gradient of temperature is non-linear and varies with both latitude and season, making it difficult for unsupervised models like SOMs, which emphasize topological smoothness, to capture abrupt transitions or fine-scale discontinuities. The scatter plot of temperature vs. depth clearly illustrates the divergence of

temperature profiles in the upper layers and across hemispheres. Furthermore, the absence of temporal features such as seasonality or day-of-year likely limited the model's ability to distinguish between warm and cold phases in similar geographic regions.

Pressure, although mechanically governed and less sensitive to dynamic oceanographic processes, showed moderate performance after being log-transformed during preprocessing. The SOM model produced a *MAE* of 0.4124 dbar, *RMSE* of 0.7324, and *MAPE* of 7.3%, but a relatively low R^2 of 0.3697. Pressure increases monotonically with depth, which should ideally make it easier to model. However, the log transformation, while stabilizing the distribution, may have flattened the depth-resolution gradient, diminishing the SOM's ability to learn depth-dependent patterns, especially in profiles with sparse sampling at greater depths. These issues were likely compounded by the underrepresentation of deep casts, as evidenced in the depth histograms, which show a strong bias toward the upper 2,000 meters.

From a modeling standpoint, the quantization error of the trained SOM was 0.0623, indicating a good fit between input vectors and their corresponding Best Matching Units. This relatively low error suggests that the SOM effectively captured the global topological structure of the dataset. However, it also indicates that most reconstruction errors arise not from poor map fitting, but from the challenge of generalizing over high-variance features, particularly temperature. In other words, while the SOM forms meaningful clusters, its capacity to reconstruct variables with sharp gradients or local discontinuities is inherently limited unless the input data provide sufficient density and diversity.

These performance differences are not random but reflect fundamental physical and statistical characteristics of each ocean variable. Salinity, which achieved the best overall performance ($R^2 \approx 0.54$), is governed by large-scale and relatively stable processes such as evaporation, precipitation, and ocean circulation, leading to smoother spatial gradients. Temperature, on the other hand, is highly sensitive to transient and small-scale drivers—such as diurnal heating and seasonal stratification—which are difficult to capture without fine temporal or vertical resolution. Pressure, despite its deterministic relationship with depth, suffered from preprocessing choices and depth sampling limitations.

In summary, the SOM model excelled at reconstructing variables with low variability and smooth gradients, like salinity, by leveraging spatial and vertical correlations across similar water masses. However, its performance was more limited for highly dynamic or structurally complex variables like temperature, where non-linear interactions and fine-scale vertical structures dominate. These results emphasize the importance of both data characteristics and model input design in driving imputation quality in oceanographic contexts.

4.2. OCEAN-SPECIFIC EVALUATION

To further investigate spatial patterns of imputation performance, the dataset was stratified by major ocean basins: Atlantic, Indian, and Pacific. Metrics were again computed only on masked values, allowing for regionally disaggregated evaluation.

In the Atlantic Ocean, salinity exhibited relatively strong performance, with *MAE* of 0.1075 PSU, *RMSE* of 0.3768, and a low *MAPE* of 0.30%. Temperature results were also slightly improved relative to the global values, with *MAE* of 1.4234°C and R^2 of 0.3322 (Table 8). The Atlantic's relatively dense observational coverage (Figure 12) may have contributed to this moderate success.

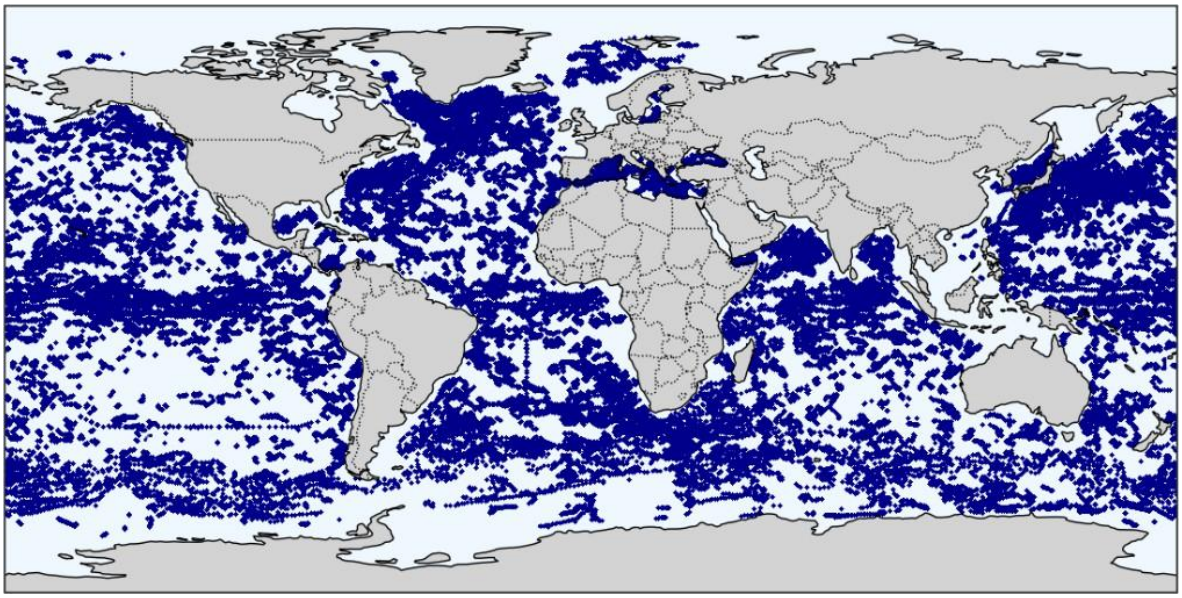


Figure 12 – Observation Coverage

The basin is well-sampled by ARGO floats and has clearer thermohaline structure, especially in the North Atlantic where water masses like North Atlantic Deep Water (NADW) and subtropical gyres generate distinct stratification. Nonetheless, the SOM still struggled to capture temperature dynamics, as seen by the *MAPE* exceeding 20%. This likely stems from the basin's sharp thermoclines and large seasonal variability in surface temperature, which are difficult to model without temporal conditioning.

Feature	MAE	RMSE	R^2	MAPE (%)
Temperature	1.4234	3.5134	0.3322	20.06
Salinity	0.1075	0.3768	0.4780	0.30
Pressure*	0.5048	0.8655	0.4466	9.96

Table 8 - SOM's Atlantic Ocean Results by Target Variable

The Indian Ocean displayed the best relative performance for temperature, with the lowest *MAE* (1.2738°C) and the highest R^2 (0.4349) among basins. Salinity also showed strong accuracy (*MAE* 0.0865 PSU; R^2 0.6854). This result may appear counterintuitive given the Indian Ocean's complex monsoonal influences, but the performance may be explained by more stable intermediate-depth profiles and predictable equatorial thermocline behavior that the SOM captured well. Additionally, the Indian Ocean's bathymetry is relatively simpler compared to the Pacific, and the ARGO float distribution, though sparser in the southern region, is fairly consistent along the equator. Pressure, however, remained modest (R^2 0.4950), suggesting limited improvement in depth-related pattern recognition despite the basin-specific focus.

Feature	MAE	RMSE	R^2	MAPE (%)
Temperature	1.2738	2.6488	0.4349	37.25
Salinity	0.0865	0.2596	0.6854	0.25
Pressure*	0.3216	0.6999	0.4950	6.22

Table 9 - SOM's Indian Ocean Results by Target Variable

In the Pacific Ocean, the SOM model exhibited its weakest performance across all variables. Temperature imputation reached a *MAE* of 1.9556°C and a *MAPE* of 28.6%, close to global averages, but with a slightly lower R^2 of 0.2658 as shown in Table 9. This is likely due to the Pacific's vast size, complex current systems (e.g., Kuroshio, Equatorial Undercurrent), and wide-ranging thermal variability from El Niño/La Niña cycles. The Pacific also includes extensive low-coverage regions and deep zones, which may have reduced the SOM's learning granularity. Salinity and pressure metrics remained comparable to other basins but with minor degradation. The low pressure R^2 (0.3166) may reflect both sampling sparsity at depth and the difficulty in capturing gradual vertical gradients over vast latitudinal spans.

Feature	MAE	RMSE	R^2	MAPE (%)
Temperature	1.9556	4.0599	0.2658	28.61
Salinity	0.0963	0.3089	0.5647	0.27
Pressure*	0.4180	0.7179	0.3166	7.16

Table 10 - SOM's Pacific Ocean Results by Target Variable

These regional results underscore that SOM-based imputation is not uniformly effective across the global ocean. Performance depends heavily on both the quality and distribution of the training data as well as the intrinsic variability of each basin. The spatial visualization confirms that the Pacific basin contains large under-sampled areas, especially near the

Southern Ocean, which are likely to produce less robust topological mappings. Additionally, regional oceanographic features such as boundary currents, eddies, and gyres introduce non-linear variability patterns that may not align well with the fixed-grid assumptions of SOMs.

4.3. SOM VS KNN: COMPARATIVE EVALUATION

The comparative analysis between the main model (SOM) and the baseline (KNN) imputation methods reveals a clear performance advantage for the KNN baseline across all variables and oceanic regions. While SOM's unsupervised architecture excels at learning latent topological structures, KNN consistently achieved lower absolute and relative error metrics (*MAE*, *RMSE*, *MAPE*) and the explanatory power (R^2).

Across global metrics, KNN achieved significantly lower errors for temperature (*MAE*: 0.3613 vs. 1.7944), salinity (*MAE*: 0.0260 vs. 0.0962), and pressure (*MAE*: 0.0927 vs. 0.4124), while also delivering much higher R^2 values (temperature: 0.91 vs. 0.29; salinity: 0.95 vs. 0.57; pressure: 0.87 vs. 0.37). These gains were not isolated—similar trends were evident in ocean-specific evaluations, where KNN demonstrated robust performance even in spatially diverse conditions such as the Pacific Ocean, where SOM exhibited weaker quantitative performance.

Several structural factors help explain these disparities:

1. **Distance-Weighted Locality vs. Topological Abstraction**

KNN directly leverages the similarity in the feature space, especially spatial proximity, by computing inverse-distance-weighted averages from the k-nearest complete observations. This mechanism inherently respects sharp local gradients and small-scale patterns, which are crucial in variables like temperature that exhibit high spatiotemporal variability. In contrast, SOM's topological mapping smooths input data onto a lower-dimensional grid, potentially leading to overgeneralization in regions where fine resolution is needed. While SOM clusters data based on similarity, it does not ensure local geometric fidelity in the same way KNN does.

2. **Curse of Dimensionality Mitigation**

In this context, the dimensionality of the data (five input features) remains relatively low, which benefits KNN. In high-dimensional settings, KNN's performance tends to deteriorate due to sparse neighborhoods, but here the feature space is compact and well-structured, allowing KNN to perform efficient, relevant imputations. SOM, on the other hand, might have been underutilized in such a low-dimensional space, where its strength in nonlinear manifold approximation is less critical.

3. Temporal Blindness of SOM

SOM did not incorporate temporal variables (e.g., day-of-year, month, or seasonal indices), which are particularly relevant for temperature imputation. Since temperature is heavily influenced by seasonal cycles, the inability of SOM to capture these patterns ay have contributed to reduced accuracy. KNN, despite also lacking explicit temporal features, indirectly benefits from spatial correlations that are often seasonally consistent due to ARGO's spatiotemporal coverage. In this way, KNN can partially compensate through dense regional profiles that act as implicit temporal proxies.

4. Imputation Strategy Design

KNN's imputation was performed using the raw distance between complete observations, inherently penalizing dissimilar profiles. This localized reasoning led to more precise reconstructions. In contrast, SOM's imputation relied on the BMU's neighborhood, which, while structured, may contain heterogeneous members—especially in areas of sparse training data. Moreover, the SOM fallback to global means when no neighbors were available likely introduced error discontinuities that KNN's fully instance-based logic avoided.

5. Training and Convergence Sensitivity

Although SOM was trained 100,000 iterations and had a quantization error of 0.062, the convergence behavior of the SOM does not guarantee optimal local resolution everywhere. The training process optimizes global topological consistency, not necessarily per-feature accuracy. KNN, being non-parametric and deterministic, avoids training altogether and guarantees consistent outputs for a given input space and neighbor configuration.

While SOM remains valuable for its interpretability, visualization capabilities, and potential for modeling global ocean patterns, this comparison highlights that for direct, fine-grained data reconstruction, especially in datasets with low-to-moderate dimensionality and rich spatial context like ARGO, KNN can be more effective in minimizing imputation error.

Nonetheless, SOM's performance should not be dismissed. Its results, though quantitatively weaker, retained coherent physical structures and could be further improved through architectural tuning or hybrid models that integrate SOM clustering with local KNN-based refinement. This direction remains a promising avenue for future development.

4.4. Limitations and Areas for Improvement

While the Self-Organizing Map approach demonstrated promising capabilities for structured imputation of ARGO profiles, several limitations emerged from both the methodology and the empirical results.

One of the clearest limitations arose in regions with sparse float coverage, where the SOM frequently failed to find populated BMU neighborhoods. In such cases, the algorithm defaulted to global mean substitution, which inevitably reduced the fidelity of the imputed values, especially for variables like pressure and temperature that vary strongly with depth and regional conditions. This fallback mechanism was most evident in areas like the Atlantic and deeper Pacific sectors, where the quantization error and lack of sufficient local analogs led to underperformance. Future implementations could incorporate adaptive fallback strategies, such as localized interpolation or hierarchical clustering models that distinguish between well-sampled and poorly sampled regions.

The exclusion of temporal variables (e.g., day-of-year, month, or season) represents another methodological gap. This design choice hindered the model's capacity to capture dynamic oceanographic phenomena such as seasonal stratification, mixed layer deepening, and thermal inertia. The observed underperformance in temperature imputation—particularly its low R^2 (0.29 globally)—likely reflects this limitation. Integrating temporal encodings or training seasonally segmented SOMs may enhance the model's sensitivity to recurring temperature structures and climatological regimes.

A single SOM was trained using the full input feature set—latitude, longitude, temperature, salinity, and pressure—which, while computationally efficient, may have introduced suboptimal trade-offs. Different variables exhibit distinct spatial and vertical behaviors: salinity showed high R^2 (approximately 0.57) and low *MAPE* due to its relatively smooth gradients, while pressure and temperature proved more challenging due to either vertical monotonicity (pressure) or high-frequency variability (temperature). Training variable-specific SOMs or implementing multi-objective learning could allow the model to better adapt to the statistical idiosyncrasies of each parameter.

Finally, the evaluation relied on a random masking strategy to simulate missing data, which does not fully reflect real-world ARGO gaps, often governed by sensor failure, depth-limited transmission, or geophysical constraints. These mechanisms introduce non-random missingness that can bias imputation if not accounted for. Future work should consider more realistic missingness simulations based on empirical ARGO float metadata or even use naturally incomplete profiles for validation. This would strengthen the external validity of the results and ensure robustness in operational applications.

By addressing these limitations—particularly through temporal enrichment, adaptive neighborhood strategies, and variable-specific modeling—future iterations of this methodology can significantly improve both the accuracy and applicability of SOM-based imputation in oceanographic data pipelines.

4.5. CONTRIBUTIONS AND IMPLICATIONS

This thesis advances the intersection of machine learning and oceanographic data processing by presenting a structured, interpretable, and computationally efficient framework for imputing missing values in ARGO float observations using Self-Organizing Maps (SOMs). The method provides spatially coherent and physically consistent reconstructions, demonstrating that unsupervised neural models can effectively address data sparsity in large-scale ocean datasets.

A key contribution lies in the model's ability to integrate geographical and environmental features—latitude, longitude, temperature, salinity, and pressure—into a unified topological structure that respects oceanographic gradients and water mass continuity. By doing so, it preserves domain-relevant relationships during imputation, offering an interpretable alternative to more opaque black-box approaches.

Beyond the ARGO dataset, this research offers a modular template for application across other ocean observing systems, such as autonomous gliders, moored instruments, and coastal sensor networks. The SOM architecture is inherently flexible and scalable, allowing for adaptation to varying spatial resolutions, sensor arrays, and platform-specific data constraints.

The broader implications of this work extend into operational oceanography and climate science. Improving the quality and continuity of in situ ocean data directly enhances the initialization of ocean models, supports more accurate short- and medium-term forecasting, and strengthens the detection of long-term climate signals. By reducing reliance on simplistic gap-filling or climatological averages, the proposed framework contributes to more robust reanalysis and informed policy-making in the context of ocean health and climate resilience.

Ultimately, this thesis highlights the strategic importance of hybrid approaches that fuse physical intuition with machine learning principles. In an era of exponentially growing ocean data, ensuring data completeness and quality will be critical. Approaches like the one developed here—grounded in both interpretability and scalability—represent a meaningful step toward enabling more reliable, data-driven ocean science.

4.6. FUTURE WORK

While the Self-Organizing Map (SOM)-based imputation framework developed in this thesis has demonstrated promising performance in reconstructing missing oceanographic variables, several avenues remain open for methodological refinement and broader application.

One of the most impactful extensions would be the explicit integration of temporal features into the model's input space. In the current approach, imputation relied solely on spatial and

environmental variables (latitude, longitude, temperature, salinity, and pressure), omitting seasonality and interannual variability. However, temperature in particular is highly sensitive to seasonal stratification, mixed layer depth, and surface forcing patterns. Incorporating temporal indicators such as month, year, or climate anomalies (e.g., ENSO phases or the Indian Ocean Dipole) could help the SOM distinguish between ocean states that are spatially similar but temporally distinct—thereby improving the model's ability to handle seasonal thermoclines and regional variability more robustly.

Another limitation lies in the deterministic nature of the imputation outputs. The current implementation produces point estimates without quantifying the uncertainty associated with each prediction. Given the sparsity and heterogeneity of ARGO data, incorporating probabilistic methods or confidence intervals would enhance the reliability of the imputations, especially in data-sparse regions or for variables with higher temporal volatility. Techniques such as Bayesian SOMs or ensemble-based SOM variants could be explored to provide not only an estimate but a distribution of plausible values.

From a methodological perspective, future work could benchmark the SOM against other unsupervised or semi-supervised models, such as Variational Autoencoders (VAEs), Generative Adversarial Imputers, or hybrid SOM–KNN architectures. These approaches might improve accuracy in high-variance scenarios or offer complementary strengths, particularly in capturing nonlinear dependencies. Although the SOM maintained spatial coherence and interpretability, KNN outperformed it in all evaluation metrics, especially for temperature and pressure. Exploring model combinations—such as using SOMs for neighborhood definition and KNN for value estimation—could reconcile the interpretability of SOM with the predictive power of distance-based methods.

The current work focused on physical variables (temperature, salinity, and pressure), but the framework can be extended to biogeochemical parameters measured by ARGO BGC floats, such as oxygen, nitrate, or pH. These variables often exhibit stronger nonlinear relationships and require variable-specific preprocessing. Their inclusion would test the adaptability of SOM architectures and the need for specialized configurations or transformations.

Finally, adapting the framework for real-time imputation scenarios represents a practical and impactful challenge. As new float data are collected and transmitted continuously, a lightweight, scalable version of the imputation model could be embedded within operational data processing pipelines. This would require automating preprocessing, adapting SOM structures incrementally, and defining fallback mechanisms for dealing with extreme sparsity or anomalous readings.

In summary, the results of this study validate SOM as a viable tool for oceanographic data imputation, particularly for preserving spatial structure and producing physically coherent estimates. However, further improvements in temporal modeling, uncertainty quantification,

model hybridity, and operational scalability will be critical in elevating this approach from a research prototype to a robust, deployable component of the ocean observing system.

5. CONCLUSION

This thesis addresses a persistent challenge in oceanographic research: how to deal with missing and irregular observational data in large-scale systems such as the ARGO program. Although ARGO represents one of the most robust and globally distributed sources of *in situ* ocean data, it suffers from inherent sparsity, uneven spatiotemporal coverage, and missing values. These limitations pose serious constraints for downstream applications, particularly ocean reanalysis, climate modeling, and long-term environmental monitoring—where data continuity and consistency are critical.

To mitigate these issues, the study proposes a Self-Organizing Map (SOM)-based imputation strategy tailored to the structure and complexity of ARGO profiles. SOMs, through unsupervised learning, are capable of clustering similar observations in a topologically coherent manner and inferring missing values based on learned patterns in both spatial and feature space. By incorporating latitude, longitude, and pressure alongside environmental variables such as temperature and salinity, the model captures geophysically meaningful structures across three dimensions.

Unlike traditional statistical approaches—such as linear interpolation or spline smoothing—which often oversimplify ocean processes and distort vertical gradients (Wong et al., 2020), the SOM preserves the ocean's physical topology without imposing artificial smoothness. This method leverages the intrinsic structure of the data, making it especially suitable for heterogeneous and nonlinear oceanographic variables.

The model's performance was evaluated via a controlled masking experiment. Overall, salinity was reconstructed with the highest accuracy ($R^2 \approx 0.54$), followed by moderate results in pressure, and lower performance in temperature. These outcomes are consistent with oceanographic principles: salinity is governed by large-scale, gradual processes—such as evaporation, precipitation, and horizontal mixing—making it more amenable to pattern-based imputation. In contrast, temperature is influenced by transient phenomena like diurnal heating, mixed layer turbulence, and seasonal thermoclines, which are harder to model without temporal inputs. Additionally, limited ARGO float density in certain regions contributed to the model's reduced performance for more dynamic variables.

Compared to supervised machine learning models such as Random Forests (Budhwani et al., 2023), MissForest (Penone et al., 2014), or Support Vector Machines (Idri et al., 2018), the SOM approach requires no complete labels, making it more practical for real-world ARGO data. While deep learning approaches like Informer-WGAN (Qian et al., 2022) have shown high accuracy on synthetic benchmarks, they demand extensive data volumes, tuning, and computational resources—factors that make their operational deployment less feasible in many scientific workflows.

This work finds its most direct parallel in Folguera et al. (2015), who applied SOMs to environmental datasets and demonstrated similar strengths in topological preservation and spatial interpretability. Like their findings, this thesis confirms that SOMs offer a balance between complexity and usability. The present approach improves upon their framework by adapting it to three-dimensional ocean profiles and evaluating it using modern performance metrics (*MAE*, *RMSE*, R^2, and *MAPE*) across ocean basins.

Notably, the SOM approach is not only technically effective but also scientifically coherent. It avoids overfitting to artificial accuracy metrics and instead reflects the physical variability and spatial structure oceanographers expect. Although the SOM did not outperform the K-Nearest Neighbors (KNN) baseline across all variables, it showed competitive performance for salinity and provided spatially and physically coherent reconstructions that KNN could not always guarantee. Its ability to encode spatial topology and profile continuity allowed it to preserve meaningful gradients—particularly for variables shaped by broad-scale processes.

Equally important is the method's interpretability, modest computational cost, and adaptability. These characteristics make it suitable for integration into operational pipelines used by organizations such as NOAA or ECMWF, where scientific transparency and reproducibility are valued alongside performance.

Still, the model has limitations. It currently lacks a temporal component, which is crucial in ocean dynamics. Its effectiveness declines in data-sparse regions, and the evaluation relied on random masking, which may not reflect structured missingness in real deployments. These shortcomings point toward future work, including seasonal or time-aware extensions, uncertainty quantification, and hybrid frameworks that couple SOM with supervised models for greater accuracy.

In conclusion, this thesis establishes a structured, interpretable, and scalable method for imputing missing values in oceanographic datasets using Self-Organizing Maps. It offers a practical and scientifically grounded tool to enhance the utility of ARGO profiles. By recovering information from incomplete records while preserving spatial and physical consistency, this method strengthens the data foundation of climate forecasting, ocean modeling, and Earth system science—domains where data fidelity shapes the trajectory of global understanding and policy.

6. BIBLIOGRAPHIC REFERENCES

Barker, P. M., Dunn, J. R., Domingues, C. M., & Wijffels, S. E. (2011). Pressure Sensor Drifts in Argo and Their Impacts. *Journal of Atmospheric and Oceanic Technology*, *28*(8), 1036–1049. https://doi.org/10.1175/2011JTECHO831.1

Budhwani, A., Lin, T., Feng, D., & Bachmann, C. (2023). Assessing and Comparing Data Imputation Techniques for Item Nonresponse in Household Travel Surveys. In *Transportation Research Record* (Vol. 2677, Issue 1, pp. 1404–1417). SAGE Publications Ltd. https://doi.org/10.1177/03611981221104802

Claustre, H., Johnson, K. S., & Takeshita, Y. (2020). Observing the Global Ocean with Biogeochemical-Argo. *Annual Review of Marine Science*, *12*(1), 23–48. https://doi.org/10.1146/annurev-marine-010419-010956

Desbruyères, D. G., Purkey, S. G., McDonagh, E. L., Johnson, G. C., & King, B. A. (2016). Deep and abyssal ocean warming from 35 years of repeat hydrography. *Geophysical Research Letters*, *43*(19), 10,356-10,365. https://doi.org/10.1002/2016GL070413

Ernst, J., & Kellis, M. (2015). Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nature Biotechnology*, *33*(4), 364–376. https://doi.org/10.1038/nbt.3157

Fazakis, N., Kostopoulos, G., Kotsiantis, S., & Mporas, I. (2020). Iterative Robust Semi-Supervised Missing Data Imputation. *IEEE Access*, *8*, 90555–90569. https://doi.org/10.1109/ACCESS.2020.2994033

Folguera, L., Zupan, J., Cicerone, D., & Magallanes, J. F. (2015). Self-organizing maps for imputation of missing data in incomplete data matrices. *Chemometrics and Intelligent Laboratory Systems*, *143*, 146–151. https://doi.org/10.1016/j.chemolab.2015.03.002

Fu, Y. B. (2014). Genetic diversity analysis of highly incomplete snp genotype data with imputations: An empirical assessment. *G3: Genes, Genomes, Genetics*, *4*(5), 891–900. https://doi.org/10.1534/g3.114.010942

Gagne, D. J., Christensen, H. M., Subramanian, A. C., & Monahan, A. H. (2020). Machine Learning for Stochastic Parameterization: Generative Adversarial Networks in the Lorenz ’96 Model. *Journal of Advances in Modeling Earth Systems*, *12*(3). https://doi.org/10.1029/2019MS001896

Grégoire, M., Garçon, V., Garcia, H., Breitburg, D., Isensee, K., Oschlies, A., Telszewski, M., Barth, A., Bittig, H. C., Carstensen, J., Carval, T., Chai, F., Chavez, F., Conley, D., Coppola, L., Crowe, S., Currie, K., Dai, M., Deflandre, B., … Yasuhara, M. (2021). A Global Ocean Oxygen Database and Atlas for Assessing and Predicting Deoxygenation and Ocean

Health in the Open and Coastal Ocean. In *Frontiers in Marine Science* (Vol. 8). Frontiers Media S.A. https://doi.org/10.3389/fmars.2021.724913

Idri, A., Abnane, I., & Abran, A. (2018). Support vector regression-based imputation in analogy-based software development effort estimation. *Journal of Software: Evolution and Process*, *30*(12). https://doi.org/10.1002/smr.2114

Jo, S. (2022). The Use of Multiple Imputation to Handle Missing Data in Secondary Datasets: Suggested Approaches when Missing Data Results from the Survey Structure. In *Inquiry (United States)* (Vol. 59). SAGE Publications Inc. https://doi.org/10.1177/00469580221088627

Johnson, G. C. (2022). Antarctic Bottom Water Warming and Circulation Slowdown in the Argentine Basin From Analyses of Deep Argo and Historical Shipboard Temperature Data. *Geophysical Research Letters*, *49*(18). https://doi.org/10.1029/2022GL100526

Johnson, G. C., Hosoda, S., Jayne, S. R., Oke, P. R., Riser, S. C., Roemmich, D., Suga, T., Thierry, V., Wijffels, S. E., & Xu, J. (2022). Argo—Two Decades: Global Oceanography, Revolutionized. *Annual Review of Marine Science*, *14*(1), 379–403. https://doi.org/10.1146/annurev-marine-022521-102008

Johnson, G. C., Purkey, S. G., Zilberman, N. V., & Roemmich, D. (2019). Deep Argo Quantifies Bottom Water Warming Rates in the Southwest Pacific Basin. *Geophysical Research Letters*, *46*(5), 2662–2669. https://doi.org/10.1029/2018GL081685

Johnson, K. S., Plant, J. N., Coletti, L. J., Jannasch, H. W., Sakamoto, C. M., Riser, S. C., Swift, D. D., Williams, N. L., Boss, E., Haëntjens, N., Talley, L. D., & Sarmiento, J. L. (2017). Biogeochemical sensor performance in the SOCCOM profiling float array. *Journal of Geophysical Research: Oceans*, *122*(8), 6416–6436. https://doi.org/10.1002/2017JC012838

Kim, T., Ko, W., & Kim, J. (2019). Analysis and Impact Evaluation of Missing Data Imputation in Day-ahead PV Generation Forecasting. *Applied Sciences*, *9*(1), 204. https://doi.org/10.3390/app9010204

Le Traon, P. Y., D'Ortenzio, F., Babin, M., Leymarie, E., Marec, C., Pouliquen, S., Thierry, V., Cabanes, C., Claustre, H., Desbruyères, D., Lacour, L., Lagunas, J. L., Maze, G., Mercier, H., Penkerc'h, C., Poffa, N., Poteau, A., Prieur, L., Racapé, V., … Xing, X. (2020). Preparing the New Phase of Argo: Scientific Achievements of the NAOS Project. *Frontiers in Marine Science*, *7*. https://doi.org/10.3389/fmars.2020.577408

Mork, K. A., Skagseth, Ø., & Søiland, H. (2019). Recent Warming and Freshening of the Norwegian Sea Observed by Argo Data. *Journal of Climate*, *32*(12), 3695–3705. https://doi.org/10.1175/JCLI-D-18-0591.1

Oke, P. R., Rykova, T., Pilo, G. S., & Lovell, J. L. (2022). Estimating Argo Float Trajectories Under Ice. *Earth and Space Science*, *9*(7). https://doi.org/10.1029/2022EA002312

Penone, C., Davidson, A. D., Shoemaker, K. T., Di Marco, M., Rondinini, C., Brooks, T. M., Young, B. E., Graham, C. H., & Costa, G. C. (2014). Imputation of missing data in life-history trait datasets: Which approach performs the best? *Methods in Ecology and Evolution*, *5*(9), 961–970. https://doi.org/10.1111/2041-210X.12232

Qian, Y., Tian, L., Zhai, B., Zhang, S., & Wu, R. (2022). Informer-WGAN: High Missing Rate Time Series Imputation Based on Adversarial Training and a Self-Attention Mechanism. *Algorithms*, *15*(7). https://doi.org/10.3390/a15070252

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). *Deep learning to represent subgrid processes in climate models*. https://doi.org/10.5281/zenodo.1402384

Roach, C. J., Balwada, D., & Speer, K. (2016). Horizontal mixing in the Southern Ocean from Argo float trajectories. *Journal of Geophysical Research: Oceans*, *121*(8), 5570–5586. https://doi.org/10.1002/2015JC011440

Roemmich, D., Alford, M. H., Claustre, H., Johnson, K. S., King, B., Moum, J., Oke, P. R., Owens, W. B., Pouliquen, S., Purkey, S., Scanderbeg, M., Suga, T., Wijffels, S. E., Zilberman, N., Bakker, D., Baringer, M. O., Belbeoch, M., Bittig, H. C., Boss, E., … Yasuda, I. (2019). On the future of Argo: A global, full-depth, multi-disciplinary array. In *Frontiers in Marine Science* (Vol. 6, Issue JUL). Frontiers Media S.A. https://doi.org/10.3389/fmars.2019.00439

Schmidtko, S., Johnson, G. C., & Lyman, J. M. (2013). MIMOC: A global monthly isopycnal upper-ocean climatology with mixed layers. *Journal of Geophysical Research: Oceans*, *118*(4), 1658–1672. https://doi.org/10.1002/jgrc.20122

Sivareddy, S., Paul, A., Sluka, T., Ravichandran, M., & Kalnay, E. (2017). The pre-Argo ocean reanalyses may be seriously affected by the spatial coverage of moored buoys. *Scientific Reports*, *7*. https://doi.org/10.1038/srep46685

Stoer, A. C., Takeshita, Y., Maurer, T. L., Begouen Demeaux, C., Bittig, H. C., Boss, E., Claustre, H., Dall'Olmo, G., Gordon, C., Greenan, B. J. W., Johnson, K. S., Organelli, E., Sauzède, R., Schmechtig, C. M., & Fennel, K. (2023). A census of quality-controlled Biogeochemical-Argo float measurements. *Frontiers in Marine Science*, *10*. https://doi.org/10.3389/fmars.2023.1233289

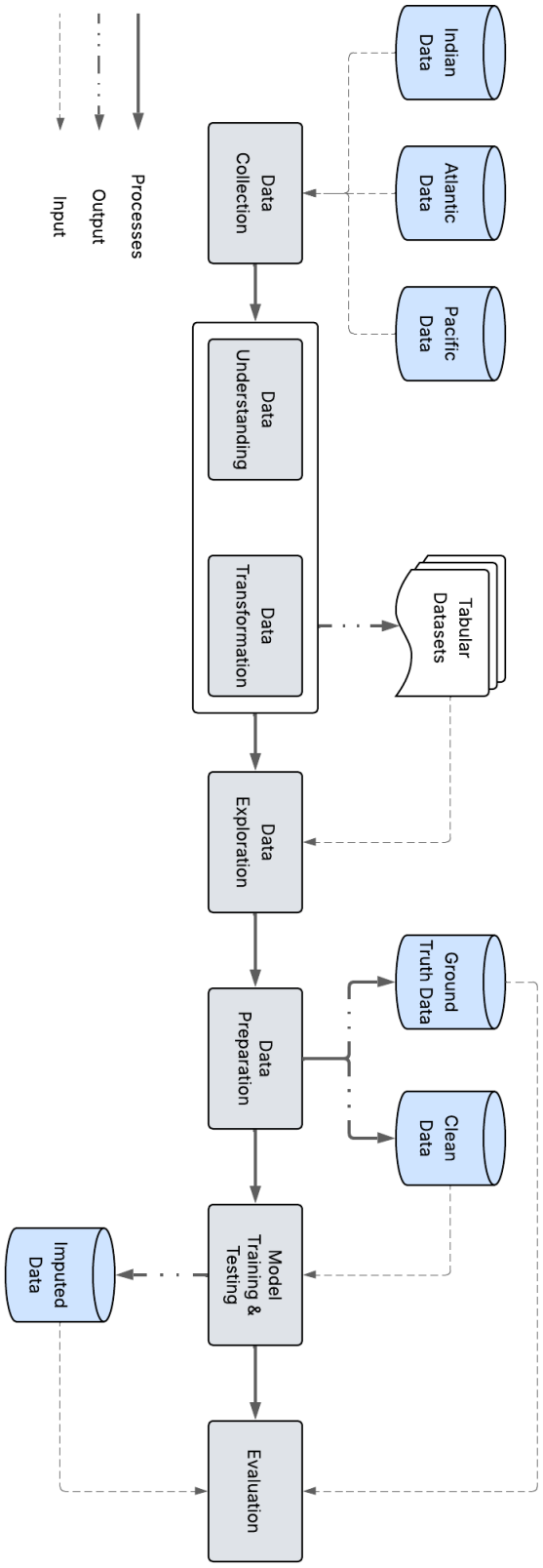
Su, H., Zhang, H., Geng, X., Qin, T., Lu, W., & Yan, X. H. (2020). OPEN: A new estimation of global ocean heat content for upper 2000 meters from remote sensing data. *Remote Sensing*, *12*(14). https://doi.org/10.3390/rs12142294

Wang, B., Fennel, K., & Yu, L. (2021). Can assimilation of satellite observations improve subsurface biological properties in a numerical model? A case study for the Gulf of Mexico. *Ocean Science*, *17*(4), 1141–1156. https://doi.org/10.5194/os-17-1141-2021

Wang, T., Wang, J., He, J., Wu, C., Luo, W., Shuai, Y., Zhang, W., & Lee, C. (2018). Investigation of the Temperature Fluctuation of Single-Phase Fluid Based Microchannel Heat Sink. *Sensors*, *18*(5), 1498. https://doi.org/10.3390/s18051498

Wong, A. P. S., Wijffels, S. E., Riser, S. C., Pouliquen, S., Hosoda, S., Roemmich, D., Gilson, J., Johnson, G. C., Martini, K., Murphy, D. J., Scanderbeg, M., Bhaskar, T. V. S. U., Buck, J. J. H., Merceur, F., Carval, T., Maze, G., Cabanes, C., André, X., Poffa, N., … Park, H. M. (2020). Argo Data 1999–2019: Two Million Temperature-Salinity Profiles and Subsurface Velocity Observations From a Global Array of Profiling Floats. In *Frontiers in Marine Science* (Vol. 7). Frontiers Media S.A. https://doi.org/10.3389/fmars.2020.00700

APPENDIX A – METHODOLOGY FRAMEWORK



APPENDIX B – DATA INGESTION VIA WEB-SCRAPPING

```
# ========================================================================

# Script Title: DATA INGESTION

# Student: Vasco Dias Pereira (20230320@novaims.unl.pt)

# Student Number: 20230320

# Date: 2024-04-22

# Description: This script is used to ingest NetCDF files.

# ========================================================================

# ========================================================================

# Title: LIBRARIES

# Description: This section imports all the required libraries for this script.

import os

import urllib.request

from bs4 import BeautifulSoup

# ========================================================================

# ========================================================================

# Title: VARIABLE ASSIGNEMENT

# Description: This section creates all the required variables used in this script.

# Data Year

dataYear = ["2019"]

# Ocean Names

oceanNames = ["atlantic", "pacific", "indian"]

# Sink File Path

sinkDirectory = "C:\\99. Pessoal\\98. Mestrado\\Thesis Code\\"

# ========================================================================

# ========================================================================
```

```
# Title: Web Scrapping Code
# Description: This section extracts all the required files.
for year in dataYear:
    for ocean in oceanNames:

        # Output folder for each year and ocean
        outputFolder = os.path.join(sinkDirectory, ocean)
        os.makedirs(outputFolder, exist_ok=True)

        # Source URL for this ocean
        sourceUrl = f"https://www.ncei.noaa.gov/data/oceans/argo/gadr/data/{ocean}/{year}"

        # Loop through each month
        for month in range(1, 13):
            monthString = f"{month:02d}"
            completeUrl = f"{sourceUrl}/{monthString}/"
            print(f"\n Accessing: {completeUrl}")

            try:
                # Reads the HTML of the current folder's month
                with urllib.request.urlopen(completeUrl) as response:
                    soup = BeautifulSoup(response.read(), 'html.parser')

                # Filters raw NetCFD files
                fileLinks = [
                    link.get('href') for link in soup.find_all('a')
```

```
            if link.get('href', '').endswith('.nc') and link.get('href', '').startswith('nodc_R')
        ]

        # Downloads each file
        for file_name in fileLinks:
            fileUrl = completeUrl + file_name
            outputPath = os.path.join(outputFolder, file_name)

            # If the file is already downloaded, ignore it
            if os.path.exists(outputPath):
                print(f"The file already exists: {file_name}")
                continue

            print(f"Downloading: {file_name}")
            urllib.request.urlretrieve(fileUrl, outputPath)

    except Exception as e:
        print(f"Error accessing {completeUrl}: {e}")
print(f"\nDownload Completed for {ocean.capitalize()}.")
# ================================================================================
```

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa