

Masters Program in **Geospatial Technologies**



CLUSTERING OCEAN STRUCTURES USING ARMOR 3D DATA FOR MARINE PATTERN RECOGNITION

Miguel Lucas

Dissertation submitted in partial fulfillment of the requirements
for the Degree of *Master of Science in Geospatial Technologies*

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

Clustering Ocean Structures Using ARMOR 3D Data for Marine Pattern Recognition

by

Miguel Filipe Baptista Lucas

Master Dissertation presented as partial requirement for obtaining the
Master's Degree in Geospatial Technologies

Supervised by

Prof. Márcia Lourenço Baptista, PhD, NOVA Information Management
School

Co-Supervised by

Prof. Marco Otávio Trindade Painho, PhD, NOVA Information
Management School
Prof. Carlos Granell Canut, PhD, Universitat Jaume I

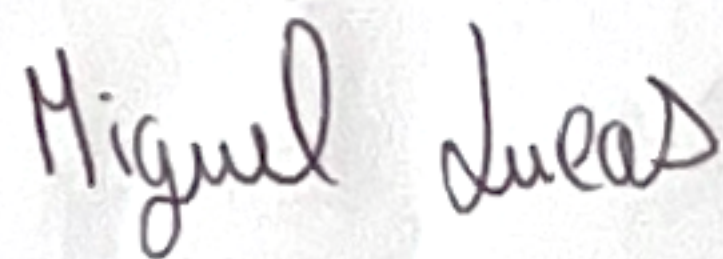
February, 2025

STATEMENT OF INTEGRITY

I declare that the work described in this document is my own and not from someone else. All the assistance I have received from other people is duly acknowledged and all the sources (published or not published) are referenced. This work has not been previously evaluated or submitted to NOVA Information Management School or elsewhere. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, 19 February 2025

Miguel Filipe Baptista Lucas

A handwritten signature in dark ink, reading "Miguel Lucas". The signature is written in a cursive style, with the first letters of "Miguel" and "Lucas" being capitalized and prominent.

USE OF GENERATIVE ARTIFICIAL INTELLIGENCE

Tasks	NO	YES	Generative Artificial Intelligence tools
Better understand issues related to the research		x	ChatGPT-4.0
Summarizing text from bibliography/resources		x	ChatGPT-4.0
Summarizing the method(s) used	x		
Translating text	x		
Grammar check		x	ChatGPT-4.0
Paraphrase or rewriting text from other people/resources		x	ChatGPT-4.0
Coding in R, Python, etc.		x	ChatGPT-4.0
Get help on a software	x		
Creating and editing images, maps, videos, etc.	x		
Data analysis	x		
Specify below other tasks not mentioned above:			

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my thesis coordinator, Prof. Márcia Lourenço Baptista, for her invaluable guidance and support throughout this research. Her insights and encouragement have been essential to my academic journey.

A special thank you to Prof. Marco Painho, the Master's coordinator, for his motivation, experience, and the knowledge he has shared, which have been instrumental in shaping my learning process.

I am also deeply grateful to all the professors from this Master's program, both from NOVA IMS and IFGI, for their dedication and for providing an exceptional academic environment.

My heartfelt appreciation goes to my parents, Ana Paula Lucas and Marco Lucas, whose incredible support and constant motivation have made this journey possible. Thank you for always ensuring that I never missed out on anything and for always believing in me.

To my colleagues from my Bachelor's degree, Carolina Cardoso, Catarina Quendera, and Rui Minez, as well as all my former professors, thank you for the knowledge, friendships, and experiences that laid the foundation for this academic path. I would also like to extend my gratitude to my closest friends, whose support and encouragement have helped me overcome challenges along the way.

A very special acknowledgment goes to my girlfriend, Carolina. Without her, I would have never joined this Master's program. Her unconditional help during the most demanding semesters and throughout the development of this thesis has been truly exceptional. I am deeply grateful for her amazing support.

Clustering Ocean Structures Using ARMOR 3D Data for Marine Pattern Recognition

Abstract

Understanding ocean structure and its variations over time is critical for climate studies, marine ecosystem research, and oceanographic monitoring. This thesis investigates the application of clustering techniques to multivariate ocean data, focusing on the ARMOR3D dataset from the Copernicus Marine Environment Monitoring Service (CMEMS). The study aims to identify patterns in ocean structures across time and depth using clustering algorithms, particularly K-means, with validation against established Ecological Marine Units (EMUs). To achieve this, the research employs a depth-stratified clustering approach, where oceanographic variables are analyzed across different depths and seasons. The clustering results from 2010 and 2022 reveal distinct seasonal and interannual changes, with surface layers exhibiting the most significant variations, while deeper ocean layers remain relatively stable. A depth-wise cluster intersection analysis ensures consistency in labeling across depth levels, facilitating a better understanding of vertical oceanographic structures. The clustering results align well with EMU classifications, supporting the effectiveness of multivariate clustering in capturing marine pattern variability. These findings contribute to advancing oceanographic data analysis by demonstrating how clustering techniques can enhance the identification and interpretation of global ocean structures.

Keywords

Multivariate Ocean Clustering, Temporal and Depth Analysis, Marine Pattern Recognition

Sustainable Development Goals (SGD):



CONTENTS

List of Figures	iv
List of Tables	v
1 Introduction	1
2 Literature Review	3
2.1 Introduction	3
2.2 3D Ocean Data	4
2.3 Dataset	5
2.4 Clustering	6
2.4.1 Overview of Clustering Models in Oceanography	6
2.4.2 Traditional Clustering Methods	8
2.4.3 Advanced and AI Driven Clustering Methods	11
2.4.4 Comparison of Methods for 3D Ocean Data	13
2.5 Case Study	15
3 Methodology	17
3.1 Data Acquisition	18
3.2 Exploratory Data Analysis	18
3.3 Data Preprocessing & Feature Engineering	19
3.4 Clustering Model & Execution	21
3.5 Temporal & Spatial Cluster Analysis	22
3.6 Validation Trough Marine Pattern Recognition	23
4 Results and Discussion	24
4.1 Spatial Distribution of Clusters	24
4.2 Mean Variable Value Per Cluster Across Depth	27
4.3 Validation Against Marine Patterns (EMU)	31
5 Conclusion and Future Research	34

LIST OF FIGURES

1	Methodology Workflow	18
2	Comparison of feature correlations between January and July 2010	19
3	Selected oceanographic variables used for clustering analysis.	20
4	Clustering Model Execution Flowchart.	22
5	Spatial distribution of clusters across depth in the months of January and July for 2010.	24
6	Spatial distribution of clusters across depth in the months of January and July for 2010.	25
7	Comparison of mean Temperature ($^{\circ}\text{C}$) per Cluster across depth in January and July for different years.	27
8	Comparison of mean Salinity (PSU) per Cluster across depth in January and July for different years.	28
9	Comparison of mean Eastward Seawater Velocity ($m\ s^{-1}$) per Cluster across depth in January and July for different years.	29
10	Comparison of mean Northward Seawater Velocity ($m\ s^{-1}$) per Cluster across depth in January and July for different years.	30

LIST OF TABLES

1	Interception of Clusters with EMU at different depths	31
---	---	----

ACRONYMS

- **ARMOR3D** – Multi Observation Global Ocean 3D Dataset
- **CMEMS** – Copernicus Marine Environment Monitoring Service
- **DBSCAN** – Density-Based Spatial Clustering of Applications with Noise
- **EDA** – Exploratory Data Analysis
- **EMU** – Ecological Marine Units
- **GOOS** – Global Ocean Observing System
- **KD-tree** – K-dimensional tree (used for spatial indexing)
- **MLD** – Mixed Layer Depth
- **NRT** – Near Real-Time Data Products
- **REP** – Reprocessed Datasets
- **SLA** – Sea Level Anomaly
- **SOM** – Self-Organizing Map
- **SST** – Sea Surface Temperature

INTRODUCTION

Research focused on ocean data analysis has come a long way since its beginning in studies about Earth's magnetic field Bauer (1920). Passing through early climate change studies like Lamb (1965) and remote sensing research focused on marine gravity anomaly like Sandwell and Smith (1997). Arriving at more complex studies like (Uppala, 2015) and Hersbach et al. (2020) that provide improved ocean wave data and enhanced spatial and temporal resolution for analyzing atmospheric and oceanic dynamics from 1950 onwards. Recent efforts, such as the work by Sayre et al. (2017), have advanced three-dimensional clustering of ocean environments. However, these studies often lack a temporal component due to the nature of the datasets used. Given that ocean structures are dynamic and evolve over time, a more comprehensive clustering framework that integrates temporal, depth-based, and spatial analysis is essential to enhance our understanding of marine pattern variability. This thesis seeks to address this gap by integrating a multivariate clustering approach that integrates both depth and time, while also validating the results using marine pattern recognition techniques. The central research question guiding this study is:

RQ: How can multivariate ocean 3D data be used to identify global ocean structures across time and depth?

To address this, the study aims to:

1. Analyze temporal variations in oceanic structures by clustering data across different years and seasons.
2. Understand how clusters are distributed along the vertical axis of the ocean, identifying how ocean structures evolve vertically.
3. Validate clustering results by comparing them with established marine ecological classifications, such as Ecological Marine Units (EMUs).

The methodology employed in this study relies on ARMOR 3D data, a globally integrated oceanographic dataset from the Copernicus Marine Environment Monitoring

Service (CMEMS). The research process begins with data acquisition, then exploratory data analysis (EDA) is conducted to examine the statistical distribution of the variables and detect potential anomalies or correlations that could influence clustering outcomes. Following this, data preprocessing and feature engineering ensure that variables are standardized, removing missing values and scaling features for consistency. The core of the analysis involves applying K-means clustering, an unsupervised learning algorithm optimized using silhouette scores to determine the optimal number of clusters at each depth level. Since oceanographic properties vary significantly with depth, clustering is performed independently for each depth layer to capture vertical structure variability. Additionally, a temporal analysis is conducted by comparing cluster distributions between two distinct years, 2010 and 2020, as well as between two seasons, January and July, to assess seasonal and interannual changes. To further refine cluster stability, a depth-wise cluster intersection analysis is implemented to ensure consistency in cluster labeling across different depth levels. Finally, the clustering results are validated using marine pattern recognition, specifically by comparing them with Ecological Marine Units (EMUs) identified in Sayre et al. (2017). This validation assesses the extent to which the detected clusters correspond to established marine ecological classifications. The clustering analysis revealed clear seasonal and interannual variations in ocean structures. Surface layers exhibited strong seasonal shifts, while deeper layers remained more stable over time. Comparing 2010 and 2022, cluster boundaries showed subtle changes, suggesting long-term oceanographic shifts. The clustering results aligned well with established Ecological Marine Units (EMUs), validating the approach. These findings demonstrate that multivariate clustering effectively captures ocean variability across time and depth, contributing to marine pattern recognition.

LITERATURE REVIEW

2.1 Introduction

Oceans are fundamental to Earth’s climate system. They help regulate global temperatures, support rich biodiversity, and influence weather patterns on a massive scale (Stommel, 1961). The study of ocean dynamics is essential to understand and manage the complex interactions between physical, chemical, and biological processes in marine environments (Pond and Pickard, 1983).

This literature review is guided by the central research question of this thesis: How can multivariate ocean 3D data be used to identify global ocean structures? Addressing this question requires a comprehensive examination of the methods used for analyzing large-scale ocean datasets.

To structure this review, we examine key areas that contribute to the understanding of oceanic clustering and its applications. First, we provide an overview of 3D ocean data. Next, we introduce the ARMOR3D dataset, detailing its structure, processing methodology, and significance for clustering applications. We then explore clustering techniques, from traditional methods to advanced AI-driven approaches. Finally, we review case studies, including the classification of Ecological Marine Units (EMUs) by Sayre et al. (2017), demonstrating how clustering reveals hidden structures in ocean data.

By reviewing these key topics, this literature review aims to provide a comprehensive synthesis of clustering applications in oceanographic studies. It will highlight the strengths and weaknesses of existing methodologies and identify research gaps that require further exploration. Ultimately, this review will serve as a foundation for applying clustering techniques to multivariate ocean data, contributing to the broader goal of improving ocean structure classification and analysis.

2.2 3D Ocean Data

Three-dimensional ocean data plays a crucial role in understanding ocean dynamics, supporting scientific research, and enabling practical applications. It is widely used in climate modeling, marine ecology, ocean circulation studies, disaster management, and resource exploration.

A possible application explored in Liu, Silver, and Bemis (2019) is mesoscale eddy tracking, which is essential to study ocean heat transport and its influence on climate change. These eddies, which transport heat and nutrients across ocean basins, require high-resolution data to be properly analyzed. Similarly, ecosystem modeling benefits from 3D data by allowing researchers to classify different water masses based on physical and chemical properties, as seen in the work in Sayre et al. (2017), which identified 37 distinct volumetric ocean regions. In the field of ocean monitoring and operational applications, digital ocean systems integrate 3D data for real-time forecasting and navigation. The China Digital Ocean Prototype System (CDOPS) developed in Zhang et al. (2011), for example, demonstrates how multi source oceanographic data can be processed to provide valuable information for marine industries. Furthermore, 3D ocean models support oil spill simulations, enhancing disaster response strategies by predicting the diffusion patterns of pollutants just as in Chen et al. (2012).

Three-dimensional ocean datasets are collected using a combination of satellite remote sensing, autonomous underwater vehicles (AUVs), buoys, ship-based surveys, and numerical models. Each method has specific advantages and challenges. Satellites such as those used in the Global Ocean Observing System (GOOS) provide extensive surface-level data but have limited penetration below the ocean surface. Buoys, drifters, and ship-based sensors offer high-resolution vertical profiles of temperature, salinity, and current velocities. For example, the World Ocean Atlas dataset integrates decades of such observations into a global ocean dataset (Sayre et al., 2017).

Once collected, ocean data goes through extensive preprocessing, cleaning, and assimilation. The Ocean Data and Information System (ODIS), for example, employs open source GIS tools to integrate and visualize large volumes of oceanographic data, ensuring it is accessible (Shesu et al., 2013).

Despite advancements, several challenges persist in 3D ocean data collection and analysis. The volume and complexity of oceanographic datasets demand significant computational resources for storage and processing (Liu, Silver, and Bemis, 2019). Furthermore, uncertainties arise due to variability in observational methods, sensor calibration issues, and gaps in spatial and temporal coverage, all of which impact data reliability (Ali et al., 2019). Visualization remains a major challenge, as traditional 2D representations struggle to fully capture the multidimensional structure of ocean features such as eddies and deep currents (Liu, Silver, and Bemis, 2019). Also, integrating multiple data sources is difficult due to variations in spatial resolutions and

temporal frequencies, complicating the fusion of information from different observational platforms (Zhang et al., 2011). These challenges highlight the need for continued advancements in data processing, quality control, and visualization techniques to enhance the usability and accuracy of 3D ocean datasets.

The integration of three-dimensional ocean data into a scientific and operational workflow has changed our understanding of ocean systems. Although challenges remain in terms of data complexity, quality control, and visualization, recent advancements in data collection and interactive visualization are enhancing the accessibility and usability of oceanographic data. Continued development in AI-driven analysis, cloud-based storage, and immersive visualization will further advance the field, making 3D ocean data an even more powerful tool for scientific research and practical applications.

2.3 Dataset

The ARMOR3D dataset (Mulet et al., 2012), (Guinehut et al., 2012) is a globally integrated multi-source oceanographic dataset developed within the Copernicus Marine Environment Monitoring Service (CMEMS). It provides Level-4 (L4) analyses of 3D ocean temperature, salinity, geopotential height, and geostrophic currents, spanning from the surface to 5500 meters depth. Additionally, it includes a 2D mixed layer depth (MLD) field, presenting insights into ocean stratification and circulation patterns.

ARMOR3D integrates both satellite-derived and in-situ ocean observations through a data assimilation process designed to maximize accuracy and spatial resolution. The dataset is structured into two primary types: Near-Real-Time (NRT) products, which offer high-frequency updates (daily and monthly), and Reprocessed (REP) datasets, which provide a long-term, high-quality time series extending from 1993 to near-present. The REP datasets undergo annual extensions to incorporate newly available data while maintaining consistency in the processing framework to prevent artificial trends or quality degradation.

Data Sources and Methodology

The ARMOR3D dataset is derived from multiple ocean observation systems, including satellite measurements of sea level anomaly (SLA), sea surface temperature (SST), and sea surface salinity (SSS), as well as in-situ profiles from the Argo program, CTD casts, and moored instruments.

The processing methodology involves the following steps:

- **Statistical Regression:** Satellite observations are projected onto vertical profiles using historical regression relationships, generating synthetic fields down to 1500m.

- **Optimal Interpolation:** These synthetic fields are merged with in-situ temperature and salinity profiles to refine the 3D representation of the ocean state.
- **Geostrophic Current Computation:** The dataset applies the **thermal wind equation** to estimate 3D geostrophic currents using ARMOR3D temperature and salinity fields along with altimeter-derived surface geostrophic velocities.

This methodology ensures that the dataset captures large-scale climatic trends and mesoscale ocean variability Mulet et al., 2012, (Guinehut et al., 2012).

Accuracy and Validation

The quality of ARMOR3D is evaluated through innovation and residual analysis, comparing modeled fields against independent in-situ data. The dataset has been found to exhibit small biases, particularly in temperature and salinity fields. For instance, temperature root-mean-square (RMS) differences are approximately 0.8°C at 100m depth and 0.2°C at 1000m, while salinity RMS differences are about 0.1 psu at 100m and 0.05 psu at 1000m. The dataset has been validated against YoMaHa Argo drift velocities for geostrophic current assessments, with zonal velocity biases around 0.06 cm/s and RMS differences of 6.4 cm/s.

A key advantage of the ARMOR3D dataset is its consistency across different time periods, ensuring that trends in oceanic properties (e.g., thermosteric sea level rise of 1 mm/year) remain free from spurious artifacts due to changes in observational coverage..

Applications in Ocean Data Clustering

The high spatial and temporal resolution of ARMOR3D makes it particularly suitable for geospatial clustering applications, such as identifying oceanic regimes based on temperature and salinity profiles, detecting mesoscale eddies, or classifying hydrographic features in multi-depth clustering frameworks. Given its multi-year coverage and improved accuracy, the dataset can be used to explore long-term oceanic patterns and their relation to climate variability, an essential aspect of advanced machine learning and clustering studies in oceanography.

2.4 Clustering

2.4.1 Overview of Clustering Models in Oceanography

Clustering models play a crucial role in oceanographic data analysis by enabling the identification of patterns and structures within complex and multidimensional datasets. The ocean is a dynamic and highly heterogeneous environment, where variations in

temperature, salinity, currents, and other factors contribute to diverse spatial and temporal structures. Clustering techniques provide a means to categorize oceanic regions based on similarities in these parameters. (Sayre et al., 2017).

Clustering models facilitate the segmentation of vast oceanographic datasets into meaningful groups, reducing data complexity and improving interpretability. This process is particularly beneficial for various applications. One such application is water mass classification, where clustering helps in identifying distinct water masses based on temperature and salinity profiles, aiding in the understanding of ocean circulation and climate patterns (Schumacher et al., 2022). Another key application is ecosystem and habitat mapping, which involves recognizing regions with similar biological or environmental characteristics. This is essential for biodiversity conservation and fisheries management according to Costello et al. (2018).

Clustering is also useful for anomaly detection, as it can help identify regions with unusual oceanographic properties, such as areas affected by climate change or pollution events (Sayre et al., 2017). Additionally, multidimensional data integration benefits from clustering methods by combining various ocean parameters across different depth levels to reveal hidden structures in the water column (Schumacher et al., 2022). Density-based clustering methods, such as the DBSCAN algorithm, are particularly useful in identifying clusters of arbitrary shapes in ocean data (Ester et al., 1996). Moreover, cluster analysis is widely used in ocean basin studies. For instance, studies like the basin-wide cluster analysis of the Atlantic Ocean provide critical insights into environmental patterns (Schumacher et al., 2022). Other advanced clustering approaches, such as hybrid K-means and ANOVA-based clustering, have been explored for improved accuracy in underwater sensor networks, offering better similarity aggregation in oceanographic data (Harb, Makhoul, and Couturier, 2015).

Given the vast amount of data collected through remote sensing, buoys, and autonomous underwater vehicles, clustering algorithms provide an efficient means of extracting knowledge without requiring predefined categories. These methods support exploratory data analysis and hypothesis generation, helping oceanographers make data-driven decisions (Jain, Murty, and Flynn, 1999).

Despite their advantages, clustering models in oceanography face several challenges. One major challenge is high dimensionality and complexity. Oceanographic data often involve multiple variables, such as temperature, salinity, and geostrophic currents, across different depths and time scales. This complexity can lead to difficulties in selecting appropriate clustering features and distance metrics (Jain, 2010). Another significant challenge is spatiotemporal variability. The dynamic nature of ocean processes introduces variability that complicates cluster stability. Seasonal changes, eddies, and other transient phenomena can affect clustering results, requiring adaptive or time-aware clustering approaches (Babanov, Semenov, and Mokhov, 2023).

The choice of the optimal number of clusters is another fundamental challenge in unsupervised learning. Traditional methods, such as the silhouette score or elbow method,

may not always yield clear results in oceanographic applications (Ester et al., 1996). Furthermore, scalability and computational costs pose difficulties. Given the size of ocean datasets, clustering methods need to be computationally efficient. Traditional clustering techniques, such as k-means, may struggle with large datasets, necessitating the use of hierarchical or deep learning-based clustering approaches (Jain, Murty, and Flynn, 1999). Recent studies suggest that enhanced K-means approaches, incorporating statistical models like ANOVA, improve clustering efficiency in oceanographic applications (Harb, Makhoul, and Couturier, 2015).

Another critical issue is the interpretability of results. While clustering models can reveal patterns, the interpretation of these clusters requires expert knowledge. Differences in parameter weighting or preprocessing steps can significantly impact results, necessitating robust validation techniques (Xu and Wunsch, 2005). The application of k-means and its variants has been widely studied, and new techniques have been developed to address cluster initialization and optimization issues, particularly in large-scale oceanographic datasets (Ikotun et al., 2023). Finally, semi-supervised clustering techniques offer a promising solution to improve clustering accuracy. Incorporating domain knowledge into clustering methods can enhance accuracy and interpretability, making semi-supervised clustering a useful compromise between unsupervised and supervised learning in oceanographic studies (Bair, 2013).

Addressing these challenges requires the continued development and adaptation of clustering algorithms to better handle ocean data's unique characteristics. Integrating clustering with machine learning, developing hybrid approaches, and leveraging high-performance computing are promising avenues for improving clustering applications in oceanography. Understanding the strengths and limitations of clustering methods will help researchers refine their analytical frameworks and contribute to more accurate ocean modeling and prediction.

2.4.2 Traditional Clustering Methods

Traditional clustering methods are fundamental techniques in data mining and analysis, widely used across various fields such as computer science, statistics, and bioinformatics. These methods aim to group objects based on their similarities across multiple dimensions, often requiring the number of clusters to be specified beforehand (Ezugwu, Ikotun, et al., 2022) (Ezugwu, Shukla, et al., 2021).

K-Means

K-means clustering (MacQueen, 1967) is one of the most widely used partitioning clustering algorithms due to its simplicity and efficiency. It has been extensively applied across various domains, including pattern recognition, data mining, and geospatial data analysis. The algorithm iteratively partitions a dataset into clusters by minimizing

intra-cluster variance, which is computed as the sum of squared distances between data points and their assigned cluster centroids (MacQueen, 1967) (Ikotun et al., 2023). The standard K-means algorithm follows a simple iterative approach:

1. Initialize k cluster centroids, either randomly or using a predefined method.
2. Assign each data point to the nearest centroid based on the Euclidean distance:

$$d(x_i, c_j) = \sqrt{\sum_{m=1}^M (x_{im} - c_{jm})^2} \quad (2.1)$$

where x_i is a data point, c_j is a cluster centroid, and M is the number of dimensions (Ahmed, Seraj, and Islam, 2020) (MacQueen, 1967).

3. Update each cluster centroid to the mean of all assigned points:

$$c_j = \frac{1}{|S_j|} \sum_{x_i \in S_j} x_i \quad (2.2)$$

where S_j is the set of points assigned to cluster j (Vincent et al., 2024).

4. Repeat steps 2 and 3 until convergence (i.e., when centroids do not change significantly between iterations).

While K-means is computationally efficient, it has several limitations. One major drawback is its dependence on the predefinition of the number of clusters, which is often unknown and requires trial-and-error or heuristic methods such as the Elbow Method or Silhouette Score for determination (Ahmed, Seraj, and Islam, 2020). Additionally, K-means is highly sensitive to initialization, meaning that poor initial centroid placement can lead to suboptimal clustering results. Methods such as K-means++ have been proposed to address this issue (Ikotun et al., 2023)(MacQueen, 1967).

Another limitation is the assumption that clusters are spherical and equally sized. K-means struggles with identifying arbitrarily shaped or density-based clusters, making it less effective in datasets where clusters have irregular structures. Alternative clustering methods, such as Kernel K-means and Fuzzy K-means, attempt to mitigate this limitation (Ahmed, Seraj, and Islam, 2020). Furthermore, K-means does not perform well in high-dimensional spaces, as the Euclidean distance measure becomes less meaningful in such cases. Distributed implementations, such as K-means using MapReduce, have been developed to improve scalability for large datasets (Ikotun et al., 2023)

K-means remains a foundational clustering algorithm despite its limitations. Advances such as improved initialization, scalability techniques, and adaptations for non-spherical clusters have enhanced its robustness. In ocean data analysis, it continues to be a valuable tool for pattern detection and environmental monitoring.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al., 1996) is a widely used clustering algorithm that identifies clusters of arbitrary shapes and isolates noisy patterns in datasets. Unlike partitioning methods such as K-means, DBSCAN does not require specifying the number of clusters in advance and is particularly effective for datasets with varying densities (Ester et al., 1996). It has been applied in numerous domains, including geospatial data analysis, pattern recognition, and big data clustering (Schubert et al., 2017).

1. **Identify core points:** A point x_i is a core point if at least MinPts points (including itself) exist within its ε -neighborhood.
2. **Expand clusters:** All directly density-reachable points from a core point are assigned to the same cluster.
3. **Identify border points:** A border point is reachable from a core point but does not have enough neighbors to be a core point itself.
4. **Identify noise:** Points that are neither core nor border points are classified as noise (Ester et al., 1996).

Mathematically, the ε -neighborhood of a point x_i is defined as:

$$N_\varepsilon(x_i) = \{x_j \mid d(x_i, x_j) \leq \varepsilon\} \quad (2.3)$$

where $d(x_i, x_j)$ represents the distance metric, typically Euclidean (Schubert et al., 2017).

A point x_i is density-reachable from x_j if:

$$\exists x_1, x_2, \dots, x_n \text{ such that } x_1 = x_j, x_n = x_i, \text{ and } x_{k+1} \in N_\varepsilon(x_k) \text{ for all } k \in \{1, \dots, n-1\} \quad (2.4)$$

(Ester et al., 1996).

Despite its advantages, DBSCAN has several limitations. The most notable drawback is its sensitivity to parameter selection, particularly epsilon and MinPts , which can significantly impact clustering results. Determining optimal values for these parameters often requires heuristic approaches or domain expertise (Weng, Gou, and Fan, 2021). Another limitation is its inefficiency on high-dimensional data. The time complexity of DBSCAN is typically $O(n^2)$ due to neighborhood queries, although spatial indexing techniques such as R-trees and kd-trees can improve performance to $O(n \log n)$ (Luchi, Rodrigues, and Varejão, 2019). Additionally, DBSCAN struggles with clusters of

varying densities and clusters with different density levels may be incorrectly merged or split (Schubert et al., 2017).

DBSCAN is a powerful clustering algorithm that remains widely used due to its ability to discover clusters of arbitrary shapes and effectively handle noise. However, its sensitivity to parameters and computational complexity in high-dimensional datasets remain challenges. Various optimizations, such as hierarchical DBSCAN (HDBSCAN) and improved neighbor search techniques, continue to enhance its applicability in real-world scenarios. In ocean data analysis, DBSCAN has demonstrated strong capabilities in detecting spatial and temporal patterns, making it a valuable tool for geospatial clustering in marine studies.

2.4.3 Advanced and AI Driven Clustering Methods

Advanced and AI-based clustering methods represent a significant evolution beyond traditional clustering techniques, leveraging artificial intelligence, deep learning, and probabilistic modeling to enhance pattern discovery and data segmentation. These methods address many limitations of classical clustering approaches by automatically determining cluster structures, handling high-dimensional data, and adapting to complex, nonlinear relationships within datasets (Ezugwu, Ikotun, et al., 2022). They have been widely applied in various domains, including geospatial analysis, bioinformatics, and computer vision (Ezugwu, Shukla, et al., 2021).

Self Organizing Maps (SOMs)

Self-Organizing Maps (SOMs), introduced by Kohonen (1982), are a type of artificial neural network used for unsupervised learning. They serve as powerful tools for clustering and data visualization by projecting high-dimensional data onto a lower-dimensional grid while preserving topological relationships (Kohonen, 1982). Due to their ability to create meaningful spatial representations of complex datasets, SOMs have been widely applied in fields such as pattern recognition, geospatial analysis, and oceanographic studies (Miljković, 2017).

The SOM algorithm consists of the following key steps:

1. **Initialization:** Assign random weights to the neurons in the map.
2. **Competition:** Identify the Best Matching Unit (BMU) for each input x_i , using the Euclidean distance:

$$d(x_i, w_j) = \sqrt{\sum_{m=1}^M (x_{im} - w_{jm})^2} \quad (2.5)$$

where w_j represents the weight vector of neuron j (Cottrell et al., 2018).

3. **Cooperation:** Adjust weights of the BMU and its neighbors using a neighborhood function $h_{j,i}$:

$$w_j(t+1) = w_j(t) + \alpha(t)h_{j,i}(t)(x_i - w_j(t)) \quad (2.6)$$

where $\alpha(t)$ is the learning rate, which decreases over time (Shieh & Liao, 2012).

4. **Iteration:** Repeat the process for a fixed number of epochs, gradually reducing the learning rate and neighborhood size.

Despite their usefulness, SOMs exhibit several limitations. One significant challenge is the selection of optimal parameters, including the size of the map, learning rate, and neighborhood function. Improper tuning can lead to poor convergence or underfitting of clusters (Wehrens and Kruisselbrink, 2018).

Another limitation is the difficulty of interpreting the resulting clusters, especially when dealing with high-dimensional data. Traditional visualization techniques such as the U-matrix help represent distances between neurons, but additional methods like ViSOM and PRSOM have been proposed to enhance interpretability (Shieh and Liao, 2012).

Moreover, SOMs assume that data are continuous and numerical, which limits their applicability to categorical datasets. Efforts have been made to extend SOMs to work with non-numerical data, such as kernel-based SOMs (Cottrell et al., 2018). Finally, computational efficiency is another concern; while batch SOMs improve performance, training large maps on massive datasets remains a challenge (Miljković, 2017).

SOMs continue to be a valuable tool for clustering and visualization in various scientific disciplines. Their ability to preserve topological structures makes them particularly useful for analyzing high-dimensional and spatial data. Despite challenges related to parameter selection, interpretability, and computational efficiency, ongoing research is addressing these issues through improved visualization techniques, hybrid models, and scalable implementations. In oceanography, SOMs remain a key method for uncovering patterns in complex marine datasets.

GeoSOM

GeoSOM (Geographic Self-Organizing Map) is an extension of the Self-Organizing Map (SOM) algorithm, specifically designed for spatial clustering. Although traditional SOM does not inherently consider the geographic context of the data, GeoSOM integrates spatial relationships to better reflect geographic dependencies (Bação, Lobo, and Painho, 2005). This adaptation makes it particularly useful for exploratory spatial data analysis, geodemographics, and environmental studies (Henriques, Bacao, and Lobo, 2012).

GeoSOM extends the standard SOM algorithm by incorporating a spatial constraint. The learning process consists of two phases:

1. **Geographic Filtering:** Instead of considering all neurons in the SOM grid, GeoSOM first restricts the search for the Best Matching Unit (BMU) to a localized geographic neighborhood. This ensures that the geographic dimension plays a central role in clustering.
2. **Attribute-Based Matching:** Within the defined geographic subset, the standard SOM distance measure is used to determine the BMU based on non-spatial attributes (Baao, Lobo, and Painho, 2005).

Mathematically, the algorithm can be described as:

$$\text{BMU} = 2 \min_{w_j \in G} d(x_i, w_j) \quad (2.7)$$

where G represents the geographically constrained set of candidate neurons, and w_j is the weight vector of neuron j (Henriques, Baao, and Lobo, 2009).

The learning update rule remains similar to SOM:

$$w_j(t + 1) = w_j(t) + \alpha(t)h_{j,i}(t)(x_i - w_j(t)) \quad (2.8)$$

where $\alpha(t)$ is the learning rate and $h_{j,i}(t)$ is the neighborhood function, which now incorporates spatial constraints (Henriques, Bacao, and Lobo, 2012).

Although GeoSOM enhances spatial clustering, it has certain limitations. One key challenge is parameter tuning, particularly defining the geographic constraint. If the constraint is too strict, important spatial patterns may be overlooked; if too loose, the results may resemble a standard SOM with reduced spatial influence (Henriques, Baao, and Lobo, 2009).

Another limitation is computational complexity. While SOM already demands significant computational resources for high-dimensional data, adding geographic constraints increases processing time. Efficient implementations, such as hierarchical GeoSOM or GPU-based acceleration, have been proposed to address this issue (Baao, Lobo, and Painho, 2005).

Furthermore, interpreting the GeoSOM results can be challenging. Unlike traditional clustering methods that provide clear partitions, GeoSOM often requires additional visualization tools, such as U-matrices and parallel coordinate plots, to extract meaningful spatial patterns (Henriques, Bacao, and Lobo, 2012).

2.4.4 Comparison of Methods for 3D Ocean Data

Clustering methods can be broadly categorized into traditional, density-based, self-organizing, and AI-based approaches. Each of these techniques has its strengths and

weaknesses, making them suitable for different types of data and analytical tasks. This section compares these clustering methods based on their algorithmic principles, computational efficiency, scalability, and applicability to geospatial and oceanographic data analysis.

Traditional clustering methods, such as K-means, rely on partitioning data into predefined clusters by minimizing intra-cluster variance. However, they assume spherical clusters and require the number of clusters to be specified beforehand, which can be a limitation (MacQueen, 1967). DBSCAN, a density-based clustering approach, overcomes this by identifying clusters of arbitrary shapes without needing the number of clusters specified, but it struggles with varying densities and high-dimensional data (Ester et al., 1996).

Self-Organizing Maps (SOMs) and their geographic variant, GeoSOM, provide an alternative by mapping high-dimensional data onto a lower-dimensional grid while preserving topological relationships. SOMs are useful for visualizing complex datasets, but their clustering performance depends on careful parameter tuning (Kohonen, 1982). GeoSOM extends this approach by explicitly incorporating spatial constraints, ensuring that neighboring data points in physical space remain close in the feature space (Bação, Lobo, and Painho, 2005). In terms of computational efficiency and scalability traditional methods like K-means are computationally efficient. In contrast, DBSCAN is more complex due to its calculation method but can be improved with spatial indexing techniques (Ester et al., 1996). On the other hand, SOMs and GeoSOMs require iterative training, making them computationally expensive for large datasets. Their training time depends on the number of neurons and input dimensions, often requiring specialized optimization techniques for large-scale applications (Bação, Lobo, and Painho, 2005). On its applicability to geospatial and ocean data analysis DBSCAN and GeoSOM offer distinct advantages. DBSCAN is well-suited for detecting oceanographic anomalies and clustering spatial data with noise, as it can separate dense regions from sparse areas (Ester et al., 1996). GeoSOM, on the other hand, explicitly considers spatial dependencies, making it a powerful tool for geodemographic studies and marine ecosystem classification (Bação, Lobo, and Painho, 2005).

Traditional K-means has been widely used in climate studies and water mass identification, though its assumptions about cluster shapes can limit its effectiveness in real-world geospatial data (MacQueen, 1967).

Each clustering method offers unique advantages and trade-offs. Traditional clustering techniques, like K-means, remain popular for their simplicity and efficiency, while DBSCAN is useful for identifying clusters of arbitrary shapes. SOMs and GeoSOM provide interpretable visualizations and spatial clustering capabilities, making them valuable for geospatial applications. The choice of method depends on the specific requirements of the dataset, including scalability, interpretability, and computational constraints.

2.5 Case Study

Case Study: Three-Dimensional Mapping of the Ocean Based on Environmental Data

A significant case study that exemplifies the application of clustering techniques in oceanographic research is the work of Sayre et al. (2017), which presents an innovative three-dimensional classification of the global ocean based on environmental parameters. This study aimed to partition the ocean into Ecological Marine Units (EMUs) using an objective, data-driven methodology, providing a global framework for understanding physical and chemical oceanographic variability.

Methodology

Sayre et al. (2017) employed a k-means clustering approach on a dataset derived from the 2013 World Ocean Atlas, which includes over 52 million ocean data points. These data points represent long-term climatological averages of six key environmental variables: temperature, salinity, dissolved oxygen, nitrate, phosphate, and silicate. The clustering process was designed to be aspatial, meaning that point location and depth were not direct determinants in the clustering algorithm. Instead, the clusters emerged solely from the statistical properties of the environmental parameters.

To determine the optimal number of clusters, the researchers utilized the pseudo-F statistic, which revealed a peak at 37 clusters, leading to the definition of 37 distinct EMUs. These EMUs were subsequently assigned depth classifications corresponding to conventional oceanographic zones: epipelagic (0–200 m), mesopelagic (200–1,000 m), bathypelagic (1,000–4,000 m), and abyssopelagic (>4,000 m). The vertical and horizontal distributions of these units provided insight into large-scale oceanic structures and hydrographic features.

Key Findings

The study identified 22 globally or regionally extensive EMUs that together accounted for 99% of the ocean's volume, while the remaining 15 were smaller and primarily found in shallow coastal regions. The clustering approach successfully delineated large-scale hydrographic features, including the Circumpolar Deep Water (CDW) in the Pacific and Indian Oceans and the North Atlantic Deep Water (NADW), validating the method's effectiveness in capturing established oceanographic patterns.

Furthermore, the study highlighted the latitudinal and vertical separation of EMUs, aligning with known biogeochemical provinces such as those described by Longhurst (2007). The results also demonstrated that environmental gradients, rather than geographic proximity, played a dominant role in structuring oceanic regions.

Relevance to Geospatial Clustering of Ocean Data

The work of Sayre et al. (2017) is highly relevant to geospatial clustering approaches in oceanography, particularly in the study of three-dimensional marine environments. It represents one of the first attempts to quantitatively partition the global ocean into volumetric regions based on environmental characteristics, providing a framework that could be further refined with additional parameters such as ocean currents, carbon flux, and biological data.

This case study demonstrates the power of clustering techniques to reveal hidden patterns in large oceanographic datasets and underscores the importance of multi-dimensional approaches when analyzing complex marine systems. Future research could extend this work by incorporating temporal variations, seasonal dynamics, and finer-resolution data to improve our understanding of marine ecosystem variability and its response to climate change.

METHODOLOGY

The objective of this study is to employ clustering algorithms to identify patterns across different seasons and years and detect known marine regions. Specifically, the analysis focuses on the ocean’s vertical structure by examining how clusters evolve across different depth layers. The foundation of this analysis is the ARMOR3D dataset from the Copernicus Marine Data Store, which provides a multi-variable, high-resolution representation of ocean conditions.

The methodology is divided in six sections as shown in Figure 1 and it begins with data acquisition, where the dataset is described, including the selected parameters and the rationale for choosing specific years and features. Following data acquisition, an exploratory data analysis (EDA) is conducted to gain insight into the dataset’s structure. After the initial analysis, the next step involves data preprocessing and feature engineering, where missing values are handled, relevant features are selected, and the data is standardized. In the fourth section the clustering process is executed using K-means. The silhouette score is used to determine the optimal number of clusters, applied independently at each depth level to account for vertical oceanographic variability. To further analyze the clustering results, in the fifth section, temporal and spatial cluster analysis is performed, focusing on seasonal and interannual changes. Graphical visualizations illustrate how cluster mean values evolve over different depth levels across seasons (January and July) and years (2010 and 2020). Additionally, a depth-wise cluster intersection analysis is conducted to ensure consistency in cluster labeling across depth layers. Finally, the clustering results are validated through marine pattern recognition, comparing the identified clusters with Ecological Marine Units (EMUs) from Sayre et al. (2017). This step provides insight into how well the clustering methodology captures established marine ecological classifications.

By following this structured methodology, the study ensures that the clustering results are scientifically robust and ecologically meaningful, allowing for a detailed examination of oceanographic patterns across both time and depth.

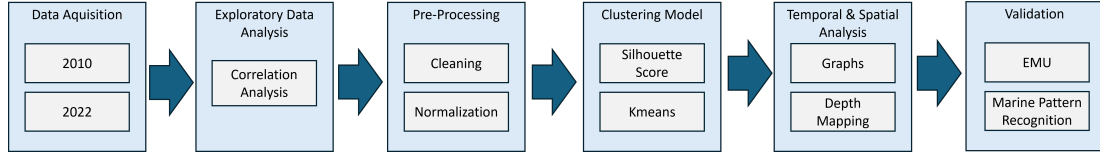


Figure 1: Methodology Workflow

3.1 Data Acquisition

The dataset used in this study was obtained from the Copernicus Marine Data Store, specifically the Multi Observation Global Ocean ARMOR3D product. This dataset provides a multidimensional representation of oceanographic parameters, incorporating observations from various sources to create a consistent and high resolution depiction of ocean conditions. The parameters represented in the dataset are: spatial coordinates (longitude and latitude), time, temperature, salinity, geostrophic currents, geopotential height and mixed-layer depth.

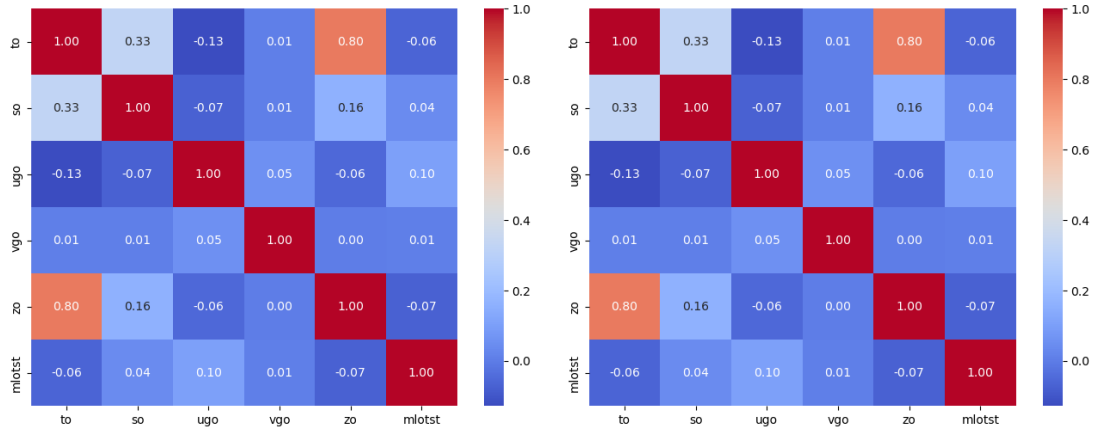
To capture temporal variations, we selected data from two distinct years, 2010 and 2020. These specific years were chosen because they were processed using the same methodology, ensuring consistency in data quality and comparability. Data from 2022 onward was excluded, as changes in processing methods introduced potential inconsistencies that could affect clustering results.

In addition, intraannual variations were considered by selecting data from two representative months: January and July. These months were chosen to explore seasonal variations, as they represent contrasting oceanographic conditions: January typically reflecting winter conditions in one hemisphere and summer in the other, while July exhibits the inverse.

By carefully selecting these temporal subsets, we aim to analyze both interannual and seasonal changes in ocean patterns, ensuring that clustering results reflect meaningful variations rather than artifacts introduced by data processing differences.

3.2 Exploratory Data Analysis

The exploratory data analysis (EDA) was conducted using Python to gain a deeper understanding of the dataset before applying clustering techniques. To investigate relationships between key oceanographic parameters, correlation matrices were generated for two representative time points, January 2010 and July 2010 (represented in Figure 2).



Feature Correlation Heatmap in January 2010 Feature Correlation Heatmap in July 2010

Figure 2: Comparison of feature correlations between January and July 2010

The correlation matrices reveal key patterns in variable dependencies. Notably, geopotential height (zo) exhibits a strong positive correlation with temperature (to) in both months, with correlation coefficients of 0.83 in January and 0.80 in July. This suggests that warmer waters are generally associated with higher sea surface height anomalies, which is expected due to the thermal expansion of water. Similarly, a moderate positive correlation is observed between salinity (so) and temperature (to), particularly in January (0.37), indicating that saltier waters tend to be found in regions of higher temperatures.

In contrast, geostrophic currents (ugo and vgo) show weak correlations with other variables, suggesting that horizontal current velocities are relatively independent of temperature, salinity, and mixed-layer depth. This aligns with ocean dynamics, where geostrophic currents are primarily driven by horizontal pressure gradients rather than local temperature or salinity changes.

3.3 Data Preprocessing & Feature Engineering

To ensure the reliability of the clustering analysis, a series of preprocessing steps were applied to clean, normalize, and refine the dataset. The first step involved data cleansing, where rows containing missing values (NaN) in key oceanographic variables were removed. This included temperature, salinity, geostrophic currents (both eastward and northward components), geopotential height (zo) and mixed-layer depth (MLD). Removing these incomplete records helped maintain data integrity and prevented missing values from introducing bias into the analysis. Notably, latitude and longitude were not considered in this filtering process, as the focus was on the oceanographic variables themselves.

Then, feature selection was performed to identify the most relevant oceanographic variables for clustering. From the available parameters: temperature (to), salinity (so), eastward and northward geostrophic currents (ugo, vgo), geopotential height (zo),

and mixed-layer depth (mldst). Only temperature, salinity, geostrophic currents, and mixed-layer depth were retained. Geopotential height (zo) was excluded from the clustering process because of the findings in 3.2. The findings from the correlation matrices highlight the importance of geopotential height (zo) as a highly correlated variable, particularly with temperature. However, it was excluded from the clustering feature set to avoid redundancy and because as it was deemed less relevant for the intended pattern detection. By selecting a subset of meaningful variables represented in Figure 3, the analysis was streamlined, reducing computational complexity while preserving essential information about oceanographic variations.

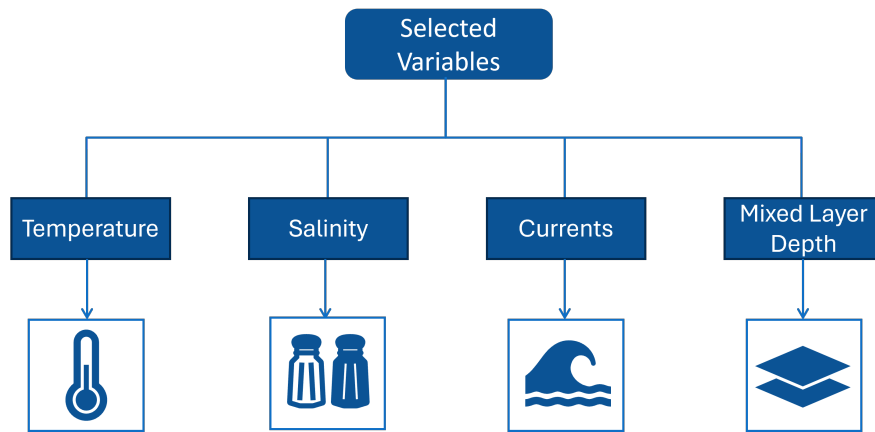


Figure 3: Selected oceanographic variables used for clustering analysis.

Following feature selection, feature scaling was applied to normalize the dataset. Given that clustering algorithms are sensitive to differences in scale, the data was standardized using the z-score transformation. This was implemented using the `StandardScaler` function from the `sklearn.preprocessing` Python package. Standardization ensures that each variable has a mean of 0 and a standard deviation of 1, making features comparable despite their differing units and ranges. This step was particularly important, as oceanographic variables, such as temperature and mixed-layer depth can have significantly different value distributions, which could otherwise distort the clustering process.

By applying these preprocessing techniques, the dataset was optimized for analysis, reducing the risk of biased results due to missing data while ensuring that all features contributed equally to the clustering model. The final dataset retained only the most relevant oceanographic parameters, ensuring a balance between data complexity and meaningful pattern discovery.

3.4 Clustering Model & Execution

The K-means clustering algorithm was selected for this study due to its computational efficiency and interpretability, aligning with the reasoning presented in the literature review. Given the large volume of oceanographic data analyzed across multiple depth levels, K-means provided a scalable solution while maintaining clear and interpretable cluster assignments. Its ability to efficiently partition data into distinct groups made it well-suited for detecting temperature and salinity anomalies, which are central to this study's objectives.

In order to determine the most suitable number of clusters, the silhouette score was used as an evaluation metric. The silhouette score measures the compactness and separation of clusters, helping to identify the most meaningful partitioning of the dataset. Since the focus of this study is on the vertical structure of the ocean, clustering was performed separately at each depth level, meaning that each depth had its own optimal number of clusters. This approach enabled a detailed exploration of how oceanographic properties evolve with depth, capturing relationships between surface conditions.

While three clusters emerged as the most frequent optimal result, it was determined that this number would not sufficiently capture the full range of oceanographic variability. Instead, the second-best clustering result of six clusters was chosen, as it provided a more nuanced representation of ocean structures while preserving interpretability. This decision allowed for a richer analysis of ocean dynamics across vertical layers.

In addition, a depth consistency approach was considered by applying the centroids from the first depth level to all depths. This process is depicted in Figure 4. The method was created to assess whether surface cluster structures persisted in deeper layers. Justifying this approach allows for an evaluation of whether deep-water characteristics are influenced by surface anomalies or if distinct clustering behavior emerges at greater depths.

By structuring the clustering process this way, the study effectively captures and interprets vertical ocean variability, ensuring that clustering results reflect meaningful seasonal and interannual changes rather than artifacts of data processing.

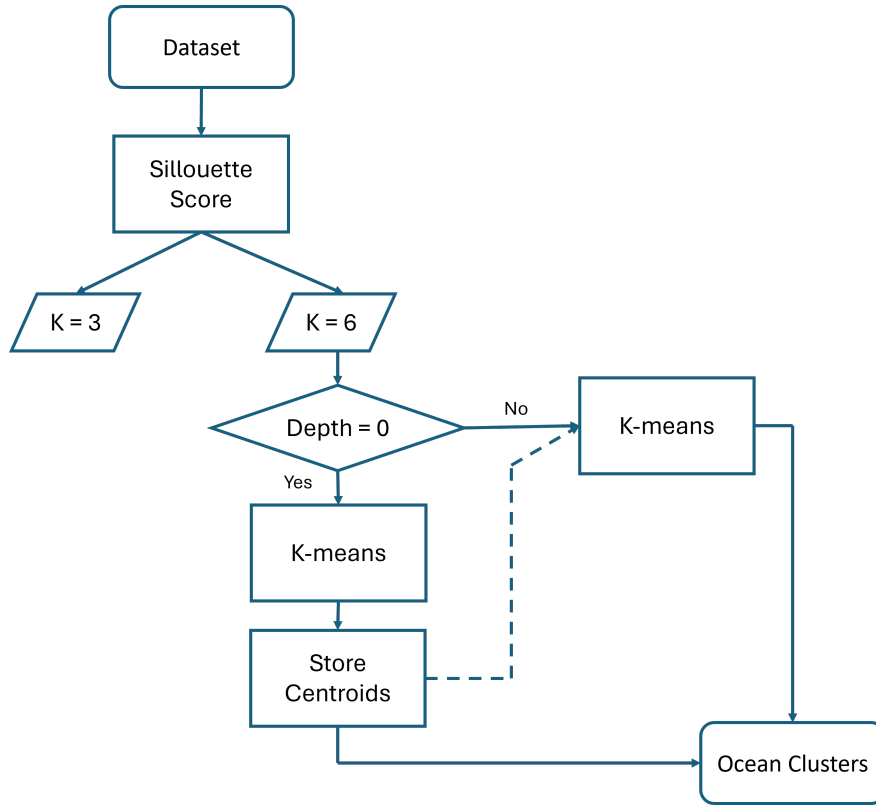


Figure 4: Clustering Model Execution Flowchart.

3.5 Temporal & Spatial Cluster Analysis

To gain a deeper understanding of oceanographic patterns, clustering results will be analyzed both seasonally and interannually through graph visualizations. These visualizations will depict the mean values of oceanographic variables for each cluster across depth, allowing an assessment of how clusters evolve over time and with depth. To evaluate seasonal changes, clustering results were compared between January and July, representing winter and summer conditions in different hemispheres.

For interannual analysis, the same methodology was applied to compare clusters between 2010 and 2020.

Since clustering was performed independently at each depth level, it was essential to ensure consistent cluster labeling across depth layers to allow vertical comparisons. To achieve this, a depth-wise cluster intersection analysis was conducted by comparing cluster assignments at consecutive depth levels. To maintain consistency, a contingency table was used to determine the best cluster match between adjacent depths. This method enabled the identification of the most similar clusters across depth layers based on their oceanographic properties, ensuring that clusters maintained coherence throughout the depth profile.

By applying these techniques, we aim to capture temporal and spatial ocean variability, ensuring that clustering results reflect meaningful patterns rather than inconsistencies

due to depth-dependent clustering assignments.

3.6 Validation Trough Marine Pattern Recognition

This validation process aims to compare the clusters derived from oceanographic datasets with the ecological marine regions identified in the study by Sayre et al. (2017). The primary objective is to assess whether the clustering methodology applied to ocean data aligns with established marine ecological classifications. To achieve this, we focus on selected depth levels (0m, 200m, 1000m, and 4000m) rather than the entire dataset. This choice follows the classical oceanographic stratification into epipelagic, mesopelagic, bathypelagic, and abyssopelagic zones, which are key in defining oceanic ecological patterns.

A spatial join was performed between the clustered ocean data and the EMUs to identify overlapping regions. Both datasets were first reprojected to a common spatial resolution to ensure comparability. Each clustered point was then assigned to the corresponding EMU based on latitude, longitude, and depth. To evaluate consistency, the proportion of clustered points falling within each EMU was computed at each depth level. This process allowed for the identification of spatial congruence between the two classifications, highlighting areas of alignment and divergence. By analyzing the percentage of intersection between the datasets.

RESULTS AND DISCUSSION

4.1 Spatial Distribution of Clusters

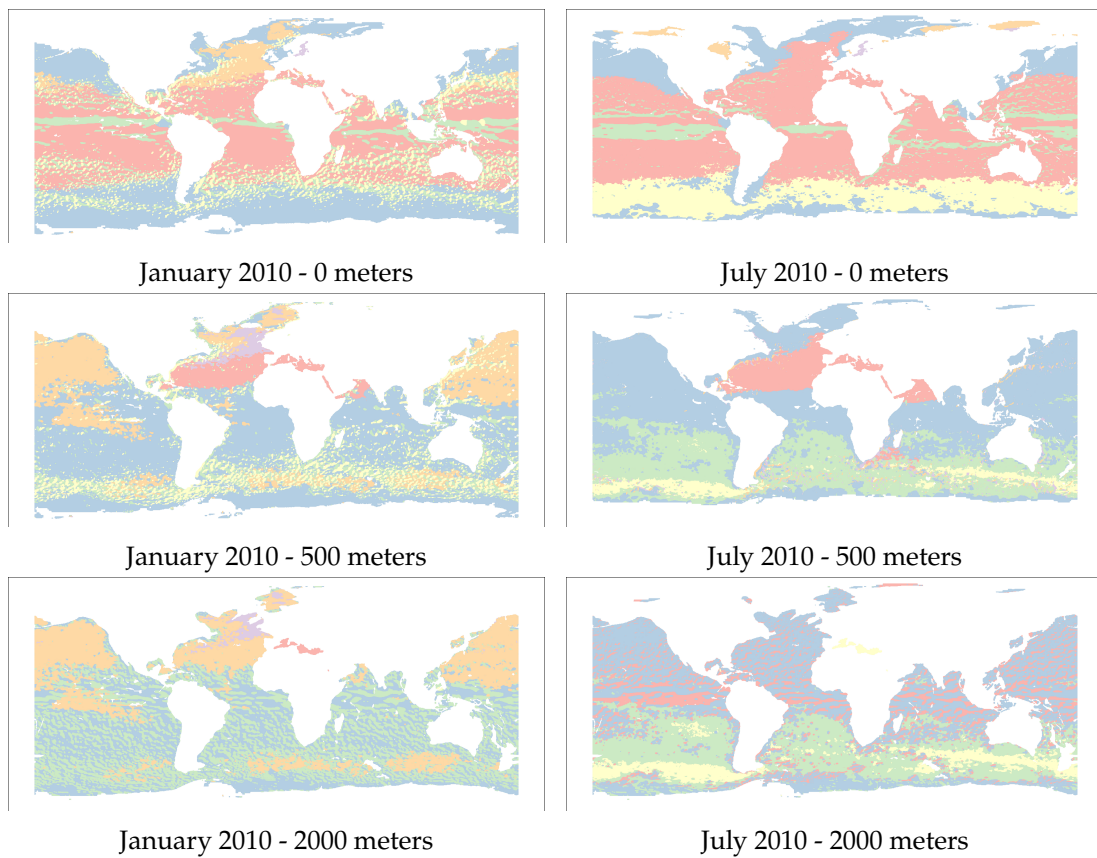


Figure 5: Spatial distribution of clusters across depth in the months of January and July for 2010.

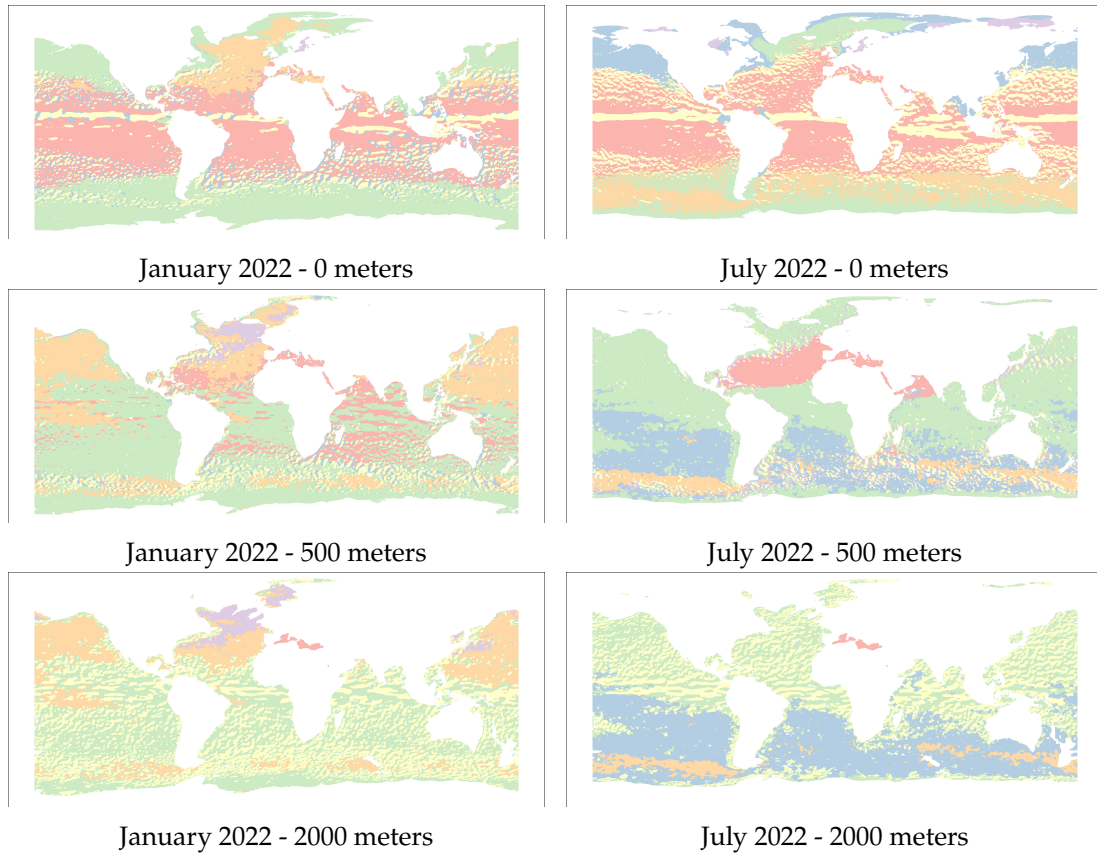


Figure 6: Spatial distribution of clusters across depth in the months of January and July for 2010.

Starting the discussion with the results spatial distribution of clusters there seems to exist patterns in the ocean that show noticeable differences between seasons and across the years 2010 and 2022. In January 2010, clustering patterns highlighted clear distinctions between polar, mid-latitude, and tropical ocean regions. The high-latitude oceans exhibited strong clustering contrasts. Tropical waters displayed relatively stable patterns, whereas mid-latitude regions showed transitions between clusters.

By July 2010, clustering patterns shifted as ocean surface layers warmed, particularly in the mid-latitude and tropical regions. Clusters in the high-latitude areas became more defined and the clustering differences between surface and deeper layers became more apparent, reflecting the summer changes.

In January 2022, clustering remained distinct across latitudes but with some variations compared to 2010. There were noticeable shifts in the distribution of clusters in polar and subpolar regions, which may indicate changes in oceanographic conditions over time. The patterns in mid-latitude and tropical regions remained largely consistent, though some adjustments in cluster boundaries were observed.

By July 2022, the clustering structure resembled previous summer patterns but with some changes in specific regions. The stratification in surface waters appeared stronger, especially in mid-latitude areas, and some cluster boundaries were altered compared

to previous years. The deep ocean showed minimal variations, consistent with the slower rate of change at those depths.

In general, clustering patterns across seasons and years highlight both seasonal cycles and potential long-term shifts in ocean conditions. While seasonal variations in cluster structures are expected, the differences between 2010 and 2022 suggest changes that could be linked to broader oceanic trends, though further analysis would be needed to understand the specific causes.

In a vertical perspective clustering patterns vary significantly with depth, with surface layers showing the most seasonal and long-term changes, while deeper layers remain more stable.

At the surface, seasonal differences are clear, with January clusters appearing more dispersed due to winter mixing and July clusters showing stronger stratification from summer warming. Between 2010 and 2022, some shifts in cluster boundaries suggest possible changes in ocean circulation or temperature patterns.

At 500m, seasonal effects are weaker, and clustering patterns remain more stable. However, over the decade, subtle differences appear, particularly in the North Atlantic and Southern Ocean, hinting at changes in intermediate water circulation.

At 2000m, clustering remains largely unchanged across seasons, reflecting the deep ocean's slow-moving nature. Minor shifts in the Atlantic and Southern Ocean between 2010 and 2022 may indicate gradual adjustments in deep-water formation.

Overall, surface layers are the most dynamic, mid-depths show some long-term trends, and deep waters remain the most stable, highlighting the vertical structure of ocean clustering over time.

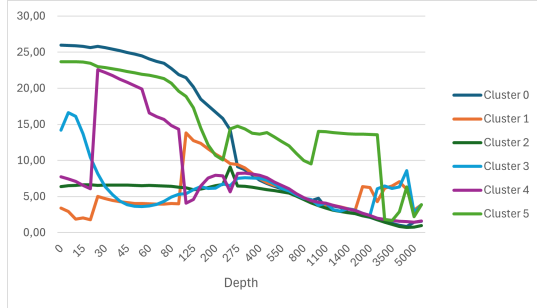
This analysis highlights how short-term variability is driven by seasonal change, while long-term trends may reflect broader shifts in ocean currents and water mass transformations. The North Atlantic and Southern Ocean stand out as key regions for long-term change, suggesting possible modifications in deep-water formation and large-scale circulation. Meanwhile, the deep ocean remains relatively stable, but slight variations indicate gradual adjustments. These findings reinforce the importance of considering both seasonal cycles and long-term trends when analyzing ocean clustering patterns across different depths.

4.2 Mean Variable Value Per Cluster Across Depth

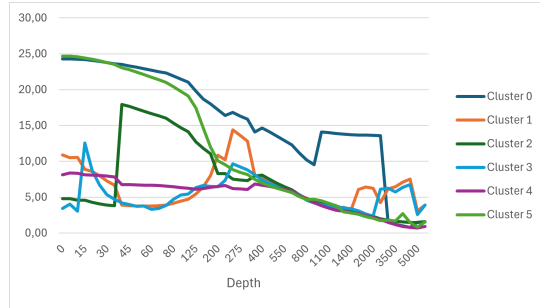
Temperature



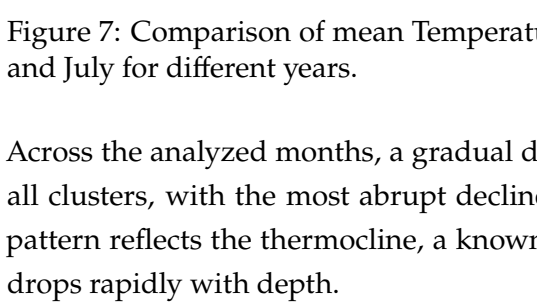
Mean Temperature (C°) per Cluster across depth in January 2010



Mean Temperature (C°) per Cluster across depth in January 2022



Mean Temperature (C°) per Cluster across depth in July 2010



Mean Temperature (C°) per Cluster across depth in July 2022

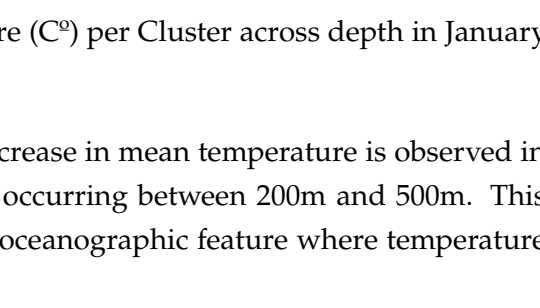


Figure 7: Comparison of mean Temperature (C°) per Cluster across depth in January and July for different years.

Across the analyzed months, a gradual decrease in mean temperature is observed in all clusters, with the most abrupt decline occurring between 200m and 500m. This pattern reflects the thermocline, a known oceanographic feature where temperature drops rapidly with depth.

Comparing January 2010 and January 2020, there is little to no significant difference in temperature distribution. Maximum values remain just below 25°C, while the lowest recorded temperature is approximately 2.5°C in the deep ocean. The clusters exhibit clear stratification in the upper ocean, remaining distinctly separated until about 200m, after which temperature values become more homogeneous and follow a slow decline. A notable exception is cluster 0, which maintains relatively high temperatures (around 12°C) between 400m and 3000m. The overall lack of interannual difference in January implies that winter temperature profiles have remained stable over the decade, with no significant warming or cooling trends detected at these depths.

In July (2010 and 2022), the general temperature structure remains similar, but stratification is less pronounced compared to January. The surface layers exhibit greater polarization in temperature values, suggesting stronger seasonal variations at shallower depths. One key seasonal difference is the depth at which the abrupt temperature drop

occurs. In January, the the decrease in values is broader and extends deeper, whereas in July, the temperature decline happens more abruptly within a narrower depth range of 200m to 300m.

Salinity

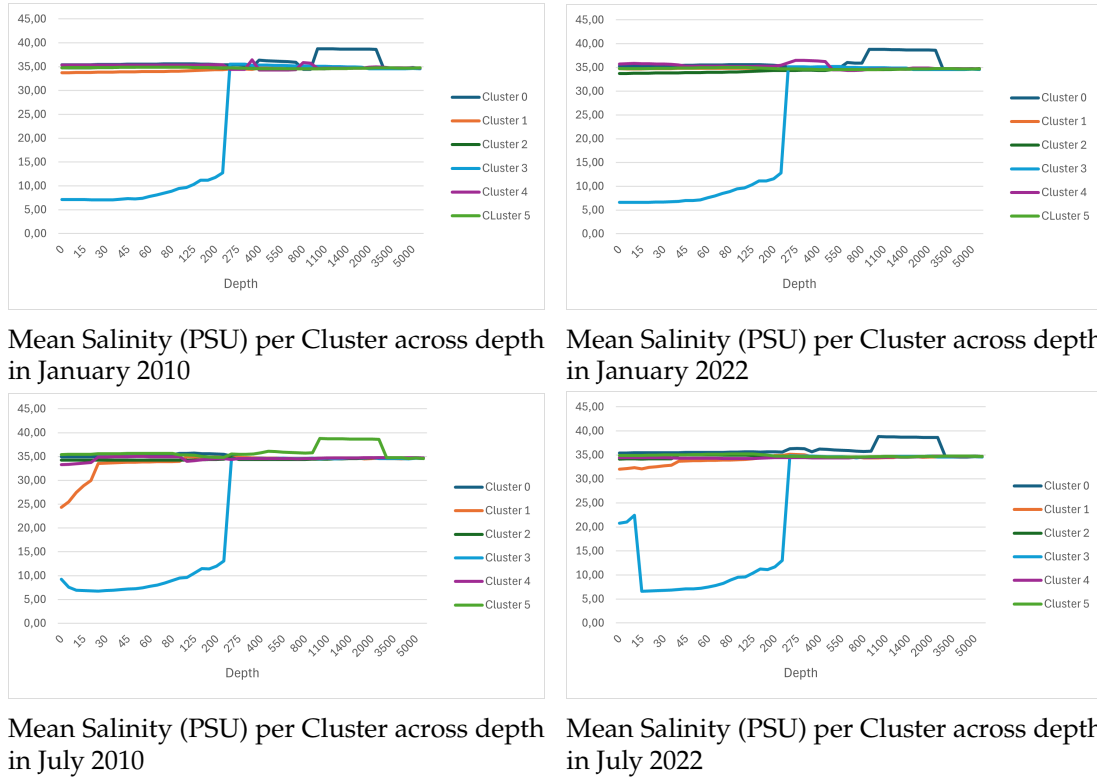


Figure 8: Comparison of mean Salinity (PSU) per Cluster across depth in January and July for different years.

When analyzing salinity patterns across depth, a striking feature is the overall consistency of most clusters, which maintain salinity values close to 35 PSU throughout the water column. However, cluster 3 deviates significantly, exhibiting a much lower salinity of approximately 7.5 PSU at the surface before converging with the other clusters around a depth of 275m. Seasonal variability is particularly noticeable in cluster 1, where a significant change occurs in July in both 2010 and 2022. In January, this cluster shows a relatively steady salinity profile, but in July, surface salinity shifts from 25 PSU to nearly 35 PSU within the first 30 meters of depth. Cluster 3 maintains a similar overall behavior in both seasons, but a key seasonal distinction emerges in the surface layer. In July 2010, surface salinity is slightly higher (10 PSU), while in July 2022, it reaches nearly 20 PSU.

Eastward Seawater Velocity



Mean Eastward Seawater Velocity ($m s^{-1}$) per Cluster across depth in January 2010

Mean Eastward Seawater Velocity ($m s^{-1}$) per Cluster across depth in January 2022



Mean Eastward Seawater Velocity ($m s^{-1}$) per Cluster across depth in July 2010

Mean Eastward Seawater Velocity ($m s^{-1}$) per Cluster across depth in July 2022

Figure 9: Comparison of mean Eastward Seawater Velocity ($m s^{-1}$) per Cluster across depth in January and July for different years.

When examining the eastward seawater velocity across depth, it is evident that variations occur primarily in deep waters, around 2000 meters and below. In the upper layers, from the surface down to approximately 2000 meters, velocity values remain close to 0 meters per second, indicating minimal horizontal movement in the eastward direction across all analyzed time frames. However, beyond this depth, there are sudden increases and decreases in velocity.

During winter months (January), cluster 3 exhibits a notable increase in velocity starting around 1400 meters, reaching 7 meters per second at 5000 meters, before experiencing a sudden drop to -7 meters per second. Interestingly, there appears to be minimal interannual change in January, implying that the general structure of these deep currents remains stable over time in the winter months.

A seasonal change is clearly observed, particularly in clusters 1, 3, and 5, which all experience sudden increases and decreases in velocity similar to what is seen in cluster 3.

Regarding interannual changes, the most significant differences appear in July, where the velocity fluctuations become more pronounced, with rises and falls reaching higher absolute values than in previous years.

Northward Seawater Velocity

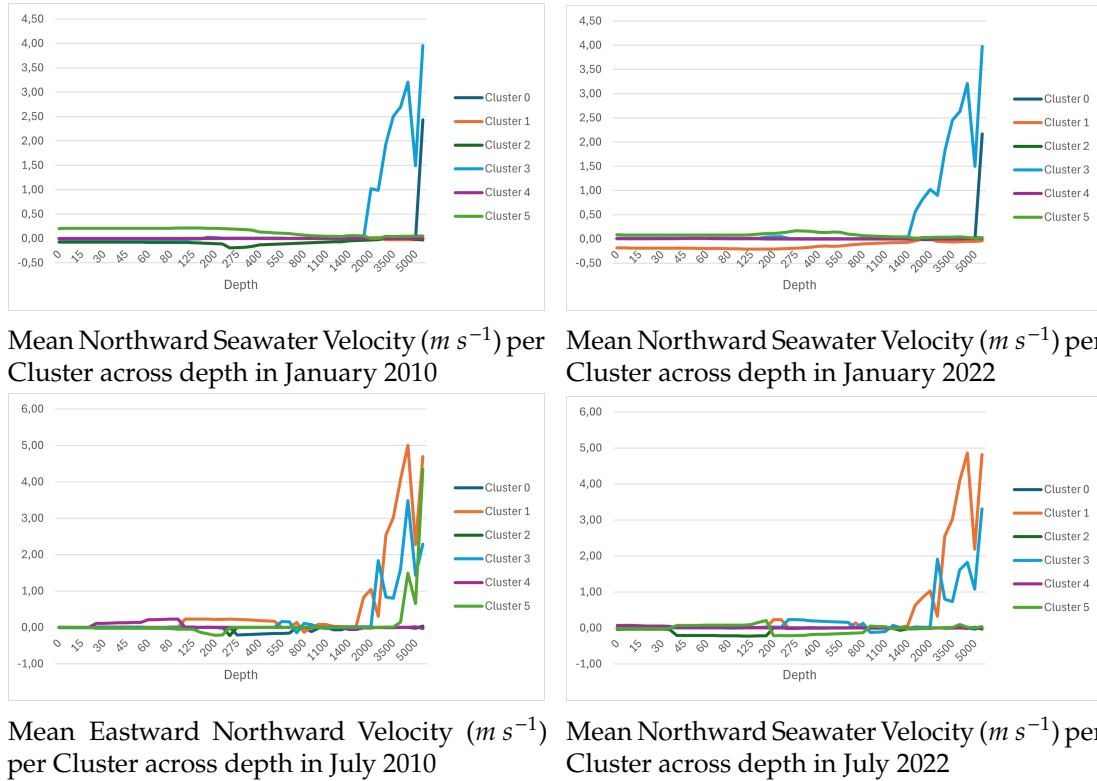


Figure 10: Comparison of mean Northward Seawater Velocity ($m s^{-1}$) per Cluster across depth in January and July for different years.

The northward seawater velocity follows a similar pattern to the eastward velocity, with variations occurring primarily in deep waters beyond 2000 meters. In all four analyzed months, there are minimal variations from 0 meters per second until 2000 meters, indicating that upper and intermediate ocean layers experience little to no significant movement in the northward direction. Unlike eastward velocity, however, there are no negative values.

In January, a significant increase in velocity is observed in Cluster 3, where there is a sudden rise to around 3 meters per second at a certain depth. Additionally, cluster 0 also experiences a sudden increase, but at a deeper level (5000 meters), reaching a slightly lower velocity of 2.5 meters per second.

When comparing seasonal variations, the overall pattern remains similar, with sudden velocity increases occurring only at deeper levels. However, in 2010, three clusters exhibited these rapid increases, whereas in 2022, only two clusters showed similar behavior. Another key seasonal difference is the increase in maximum velocity values during July. While the absolute peak value remains at 5 meters per second, the majority of the clusters that experience sudden increases reach around 3 meters per second, which is higher than in January.

When examining interannual changes, the most noticeable trend is the reduction in

the number of clusters experiencing sudden velocity increases from three clusters in 2010 to only two clusters in 2022. Moreover, while the overall behavior remains stable, the magnitude of velocity increases is slightly higher in July 2022 compared to January 2010.

4.3 Validation Against Marine Patterns (EMU)

Table 1: Interception of Clusters with EMU at different depths

Depth	Cluster	Name Emu	Percentage (%)
0	0	29	74.16
	1	3	42.94
	2	29	67.19
	3	22	60.37
	3	23	39.42
	4	6	26.16
	4	18	25.58
	4	29	26.70
	5	29	46.50
200	0	29	77.07
	1	3	43.31
	2	29	37.71
	3	22	88.89
	4	6	27.35
	4	18	26.23
	5	29	28.02
1000	0	30	85.50
	1	29	50.32
	2	29	37.86
	3	6	53.32
	4	29	49.69
	5	29	43.75
4000	0	29	66.42
	1	29	40.04
	2	29	62.81
	3	29	100.00
	4	29	56.53
	5	29	42.65

Moving the discussion into the validation process, based on the results provided in 1 in the surface level (0m), there seems to exist is a strong agreement between the clusters and the ecological marine units (EMUs). For example, cluster 0 aligns with EMU 29 at 74.16% in one dataset and increases to 77.07% in another. In a separate dataset, cluster 0 is associated with EMU 30 at 85.50%, reinforcing the idea that surface ocean regions

are well defined. At 200m, the clustering method still shows a strong alignment with the EMUs, though with some variation. For instance, cluster 2 overlaps with EMU 29 at 67.19% in one dataset but drops to 37.71% in another. Similarly, cluster 3 aligns with EMU 22 at 60.37%, yet in a different dataset, this match increases to 88.89%. By 1000m (bathypelagic zone), the alignment between clustering and the EMUs begins to weaken. In one case, cluster 3 intersects with EMU 6 at 53.32%, showing a moderate level of agreement. Cluster 4, which had stronger alignment at shallower depths, now overlaps with EMU 29 at only 49.69%, and in another dataset, this drops further to 26.16%. At 4000m, the results present an interesting contrast. One dataset shows cluster 3 aligning with EMU 29 at 100.00%, meaning the clustering method perfectly captured a distinct deep-sea region. However, other clusters show more variability. For instance, cluster 5 intersects with EMU 29 at only 42.65%, while cluster 4 shows a 56.53% match with the same EMU.

Overall, the clustering method performs best at shallower depths, where oceanographic structures are more dynamic and more distinct. The strongest agreement occurs at 0m, where surface currents and temperature gradients create clear ecological zones, followed by 200m, where clustering still aligns well with EMUs despite some regional variations. At 1000m, the clustering method begins to struggle, reflecting the increasing complexity of deep-sea circulation and the lack of sharp ecological boundaries. Finally, at 4000m, the clustering approach produces mixed results, with one perfect match but other clusters showing weaker alignment with predefined EMUs.

These findings suggest that clustering is an effective tool for detecting major ocean regions, but its reliability decreases with depth due to the increasing uniformity and complexity in deeper ocean environments.

This study investigated how multivariate ocean 3D data can be used to identify global ocean structures across time and depth through clustering. The results demonstrate that clustering effectively captures spatial and temporal patterns in oceanographic data, with clear seasonal and inter annual variations, as well as distinct vertical structures. Seasonal changes were evident across all depth levels, with surface waters showing the strongest variations. Winter months exhibited more dispersed clustering patterns due to enhanced mixing, while summer months displayed stronger stratification, particularly in mid-latitude regions. Comparing 2010 and 2022, clustering patterns remained largely consistent, but subtle shifts were observed, particularly in high-latitude regions, suggesting potential long-term changes in ocean circulation and water mass distribution.

Vertically, clustering structures varied significantly with depth. Surface layers exhibited the highest variability due to seasonal and inter annual influences, while deeper layers remained relatively stable. The mid level depths emerged as a critical transition zone where rapid changes in temperature and salinity influenced cluster formation. Below 1000m, clustering patterns showed only minor variations over time, consistent with

the slow-changing nature of deep-sea circulation. However, slight shifts detected in the North Atlantic and Southern Ocean suggest gradual adjustments in deep-water formation processes.

Validation against Ecological Marine Units (EMUs) confirmed that clustering aligns well with known oceanographic structures, particularly at shallower depths. At the surface levels, strong overlaps with EMU regions indicate that clustering effectively captures key marine patterns. However, at deeper depths, the alignment weakened, reflecting the increasing uniformity of the deep ocean environments.

These findings answer the research question: How can multivariate ocean 3D data be used to identify global ocean structures across time and depth? Clustering proves to be a valuable method for detecting both seasonal and long-term variations in ocean structure. It effectively identifies well-defined regions at the surface and mid-depths, capturing both dynamic seasonal transitions and more gradual interannual shifts. However, its applicability decreases with depth due to the lack of strong physical gradients and the complex nature of deep-sea circulation. While clustering successfully delineates major oceanic structures, further refinements are needed to enhance its performance in deeper environments.

CONCLUSION AND FUTURE RESEARCH

This study aimed to explore the application of clustering techniques to identify global ocean structures across time and depth using ARMOR3D oceanographic data. The central research question focused on how multivariate ocean 3D data can be utilized to detect and classify dynamic marine patterns. To address this, clustering was performed on ocean temperature, salinity, geostrophic currents, and mixed-layer depth, with particular emphasis on temporal and vertical variability.

The results demonstrated that clustering successfully captured significant seasonal and interannual variations in oceanographic patterns. Seasonal shifts were particularly evident in surface layers, while deeper ocean structures exhibited more stability. Additionally, differences between 2010 and 2022 indicated potential long-term changes in ocean circulation, which could be linked to broader climatic trends. The validation process comparing clusters with Ecological Marine Units (EMUs) provided further insights into the relevance of the identified clusters, showing a substantial overlap with known marine ecological regions.

Despite the success of the clustering methodology in detecting spatial and temporal ocean patterns, several limitations must be acknowledged. The depth-wise clustering approach, while effective in capturing vertical variations, may have overlooked interactions between depth layers. A more integrated approach that incorporates depth as a feature within the clustering process could produce a more holistic understanding of ocean structures. Additionally, the choice of K-means clustering, although computationally efficient and widely used, assumes spherical clusters and equal variance, which may not fully capture the complexity of oceanic features. Alternative methods, such as DBSCAN or hierarchical clustering, may provide more flexibility in detecting irregular and density-dependent structures. Furthermore, the study's reliance on ARMOR3D data from only two years (2010 and 2022) and two representative months (January and July) limits the ability to identify longer-term oceanic trends. A more extensive temporal dataset spanning multiple decades would allow for a deeper exploration of ocean variability and its potential links to climate change.

Future research should focus on alternative clustering techniques and extended temporal analysis to further refine ocean pattern recognition. Exploring density-based methods such as DBSCAN and AI-driven clustering approaches like Self-Organizing Maps (SOM) or deep learning-based clustering could improve the detection of complex and nonlinear ocean structures. These methods may provide better adaptability to heterogeneous marine environments where K-means assumptions may not hold. Additionally, expanding the dataset to cover multiple decades would allow for a more thorough investigation of long-term oceanic variability and potential climate change effects. A broader temporal perspective could enhance the detection of slow-moving patterns and improve the robustness of clustering results.

This research contributes to the growing field of geospatial ocean pattern recognition by demonstrating the potential of clustering techniques in identifying dynamic marine structures across depth and time. The study confirms that clustering can reveal significant spatial, seasonal, and interannual patterns in ocean data, emphasizing the need for continued advancements in multivariate marine analysis. As ocean datasets grow and machine learning methodologies evolve, further refinement of clustering techniques will enhance our ability to detect and predict critical oceanic changes, supporting marine conservation and climate research efforts.

BIBLIOGRAPHY

- Ahmed, M., R. Seraj, and S. M. S. Islam (2020). "The k-means algorithm: A comprehensive survey and performance evaluation". In: *Electronics* 9.8, p. 1295. DOI: [10.3390/electronics9081295](https://doi.org/10.3390/electronics9081295) (cit. on p. 9).
- Ali, W. H. et al. (2019). "SeaVizKit: Interactive Maps for Ocean Visualization". In: *OCEANS 2019 MTS/IEEE SEATTLE*. IEEE, pp. 1–10. DOI: [10.23919/OCEANS40490.2019.8962794](https://doi.org/10.23919/OCEANS40490.2019.8962794) (cit. on p. 4).
- Babanov, B. A., V. A. Semenov, and I. I. Mokhov (2023). "Comparison of Cluster Analysis Methods for Identifying Weather Regimes in the Euro-Atlantic Region for Winter and Summer Seasons". In: *Izvestiya, Atmospheric and Oceanic Physics* 59.6, pp. 605–623. DOI: [10.1134/S0001433823060026](https://doi.org/10.1134/S0001433823060026) (cit. on p. 7).
- Baço, F., V. Lobo, and M. Painho (2005). "The self-organizing map, the Geo-SOM, and relevant variants for geosciences". In: *Computers & Geosciences* 31.2, pp. 155–163. DOI: [10.1016/j.cageo.2004.06.013](https://doi.org/10.1016/j.cageo.2004.06.013) (cit. on pp. 12–14).
- Bair, E. (2013). "Semi-supervised clustering methods". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 5.5, pp. 349–361. DOI: [10.1002/wics.1270](https://doi.org/10.1002/wics.1270) (cit. on p. 8).
- Bauer, L. A. (1920). "Some of the chief problems in terrestrial magnetism and electricity". In: *Proceedings of the National Academy of Sciences* 6.10, pp. 572–580. DOI: [10.1073/pnas.6.10.572](https://doi.org/10.1073/pnas.6.10.572) (cit. on p. 1).
- Chen, G. et al. (2012). "Design and Implementation of a 3D Ocean Virtual Reality and Visualization Engine". In: *Journal of Ocean University of China* 11, pp. 481–487. DOI: [10.1007/s11802-012-2112-6](https://doi.org/10.1007/s11802-012-2112-6) (cit. on p. 4).
- Costello, M. J. et al. (2018). "Stratifying ocean sampling globally and with depth to account for environmental variability". In: *Scientific Reports* 8.1, pp. 1–9. DOI: [10.1038/s41598-018-21504-8](https://doi.org/10.1038/s41598-018-21504-8) (cit. on p. 7).
- Cottrell, M. et al. (2018). "Self-organizing maps, theory and applications". In: *Revista de Investigacion Operacional* 39.1, pp. 1–22. DOI: [10.13140/RG.2.2.14202.08648](https://doi.org/10.13140/RG.2.2.14202.08648) (cit. on p. 12).

- Ester, M. et al. (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise". In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*. AAAI Press, pp. 226–231 (cit. on pp. 7, 8, 10, 14).
- Ezugwu, A. E., A. M. Ikotun, et al. (2022). "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects". In: *Engineering Applications of Artificial Intelligence* 110, p. 104743. DOI: [10.1016/j.engappai.2022.104743](https://doi.org/10.1016/j.engappai.2022.104743) (cit. on pp. 8, 11).
- Ezugwu, A. E., A. K. Shukla, et al. (2021). "Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature". In: *Neural Computing and Applications* 33, pp. 6247–6306. DOI: [10.1007/s00521-020-05583-6](https://doi.org/10.1007/s00521-020-05583-6) (cit. on pp. 8, 11).
- Guinehut, S. et al. (2012). "High resolution 3-D temperature and salinity fields derived from in situ and satellite observations". In: *Ocean Science* 8.5, pp. 845–857. DOI: [10.5194/os-8-845-2012](https://doi.org/10.5194/os-8-845-2012) (cit. on pp. 5, 6).
- Harb, H., A. Makhoul, and R. Couturier (2015). "An enhanced K-means and ANOVA-based clustering approach for similarity aggregation in underwater wireless sensor networks". In: *IEEE Sensors Journal* 15.10, pp. 5483–5493. DOI: [10.1109/JSEN.2015.2421871](https://doi.org/10.1109/JSEN.2015.2421871) (cit. on pp. 7, 8).
- Henriques, R., F. Bacao, and V. Lobo (2012). "Exploratory geospatial data analysis using the GeoSOM suite". In: *Computers, Environment and Urban Systems* 36.3, pp. 218–232. DOI: [10.1016/j.compenvurbsys.2011.11.003](https://doi.org/10.1016/j.compenvurbsys.2011.11.003) (cit. on pp. 12, 13).
- Henriques, R., F. Bação, and V. Lobo (2009). "GeoSOM Suite: A tool for spatial clustering". In: *Computational Science and Its Applications–ICCSA 2009: International Conference, Seoul, Korea, June 29–July 2, 2009, Proceedings, Part I*. Springer, pp. 453–466. DOI: [10.1007/978-3-642-02454-2_32](https://doi.org/10.1007/978-3-642-02454-2_32) (cit. on p. 13).
- Hersbach, H. et al. (2020). "The ERA5 global reanalysis". In: *Quarterly Journal of the Royal Meteorological Society* 146.730, pp. 1999–2049. DOI: [10.1002/qj.3803](https://doi.org/10.1002/qj.3803) (cit. on p. 1).
- Ikotun, A. M. et al. (2023). "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data". In: *Information Sciences* 622, pp. 178–210. DOI: [10.1016/j.ins.2022.11.139](https://doi.org/10.1016/j.ins.2022.11.139) (cit. on pp. 8, 9).
- Jain, A. K. (2010). "Data clustering: 50 years beyond K-means". In: *Pattern Recognition Letters* 31.8, pp. 651–666. DOI: [10.1016/j.patrec.2009.09.011](https://doi.org/10.1016/j.patrec.2009.09.011) (cit. on p. 7).
- Jain, A. K., M. N. Murty, and P. J. Flynn (1999). "Data Clustering: A Review". In: *ACM Computing Surveys (CSUR)* 31.3, pp. 264–323. DOI: [10.1145/331499.331504](https://doi.org/10.1145/331499.331504) (cit. on pp. 7, 8).
- Kohonen, T. (1982). "Self-organized formation of topologically correct feature maps". In: *Biological Cybernetics* 43.1, pp. 59–69. DOI: [10.1007/BF00337288](https://doi.org/10.1007/BF00337288) (cit. on pp. 11, 14).

- Lamb, H. H. (1965). "The early medieval warm epoch and its sequel". In: *Palaeogeography, Palaeoclimatology, Palaeoecology* 1, pp. 13–37. DOI: [10.1016/0031-0182\(65\)90004-0](https://doi.org/10.1016/0031-0182(65)90004-0) (cit. on p. 1).
- Liu, L., D. Silver, and K. Bemis (2019). "Visualizing Three-Dimensional Ocean Eddies in Web Browsers". In: *IEEE Access* 7, pp. 44734–44747. DOI: [10.1109/ACCESS.2019.2909655](https://doi.org/10.1109/ACCESS.2019.2909655) (cit. on p. 4).
- Luchi, D., A. L. Rodrigues, and F. M. Varejão (2019). "Sampling approaches for applying DBSCAN to large datasets". In: *Pattern Recognition Letters* 117, pp. 90–96. DOI: [10.1016/j.patrec.2018.11.018](https://doi.org/10.1016/j.patrec.2018.11.018) (cit. on p. 10).
- MacQueen, J. (1967). "Some methods for classification and analysis of multivariate observations". In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 1. University of California Press, pp. 281–297. DOI: [10.1525/9780520313880-018](https://doi.org/10.1525/9780520313880-018) (cit. on pp. 8, 9, 14).
- Miljković, D. (2017). "Brief review of self-organizing maps". In: *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, pp. 1061–1066. DOI: [10.23919/MIPRO.2017.7973620](https://doi.org/10.23919/MIPRO.2017.7973620) (cit. on pp. 11, 12).
- Mulet, S. et al. (2012). "A new estimate of the global 3D geostrophic ocean circulation based on satellite data and in-situ measurements". In: *Deep Sea Research Part II: Topical Studies in Oceanography* 77, pp. 70–81. DOI: [10.1016/j.dsr2.2012.04.012](https://doi.org/10.1016/j.dsr2.2012.04.012) (cit. on pp. 5, 6).
- Pond, S. and G. L. Pickard (1983). *Introductory Dynamical Oceanography*. Pergamon Press Ltd. DOI: [10.1016/B978-0-08-057054-1.50001-2](https://doi.org/10.1016/B978-0-08-057054-1.50001-2) (cit. on p. 3).
- Sandwell, D. T. and W. H. Smith (1997). "Marine gravity anomaly from Geosat and ERS 1 satellite altimetry". In: *Journal of Geophysical Research: Solid Earth* 102.B5, pp. 10039–10054. DOI: [10.1029/96JB03223](https://doi.org/10.1029/96JB03223) (cit. on p. 1).
- Sayre, R. G. et al. (2017). "A Three-Dimensional Mapping of the Ocean Based on Environmental Data". In: *Oceanography* 30.1, pp. 90–103. DOI: [10.5670/oceanog.2017.116](https://doi.org/10.5670/oceanog.2017.116) (cit. on pp. 1, 4, 7, 15, 23).
- Schubert, E. et al. (2017). "DBSCAN revisited, revisited: why and how you should (still) use DBSCAN". In: *ACM Transactions on Database Systems (TODS)* 42.3, pp. 1–21. DOI: [10.1145/3068335](https://doi.org/10.1145/3068335) (cit. on p. 11).
- Schumacher, M. et al. (2022). "The Atlantic Ocean landscape: A basin-wide cluster analysis of the Atlantic near seafloor environment". In: *Frontiers in Marine Science* 9, p. 936095. DOI: [10.3389/fmars.2022.936095](https://doi.org/10.3389/fmars.2022.936095) (cit. on p. 7).
- Shesu, R. V. et al. (2013). "Open Source Architecture for Web-Based Oceanographic Data Services". In: *Data Science Journal* 12, pp. 47–55. DOI: [10.2481/dsj.13-005](https://doi.org/10.2481/dsj.13-005) (cit. on p. 4).
- Shieh, S.-L. and I.-E. Liao (2012). "A new approach for data clustering and visualization using self-organizing maps". In: *Expert Systems with Applications* 39.15, pp. 11924–11933. DOI: [10.1016/j.eswa.2012.02.162](https://doi.org/10.1016/j.eswa.2012.02.162) (cit. on p. 12).

- Stommel, H. (1961). "Thermohaline Convection with Two Stable Regimes of Flow". In: *Tellus* 13.2, pp. 224–230. DOI: [10.1111/j.2153-3490.1961.tb00079.x](https://doi.org/10.1111/j.2153-3490.1961.tb00079.x) (cit. on p. 3).
- Vincent, A. N. et al. (2024). "Multivariate Analysis of Clustering Algorithms for Marine Data". In: *2024 3rd International Conference on Sentiment Analysis and Deep Learning (ICSADL)*. IEEE, pp. 31–38. DOI: [10.1109/ICSADL61749.2024.00011](https://doi.org/10.1109/ICSADL61749.2024.00011) (cit. on p. 9).
- Wehrens, R. and J. Kruisselbrink (2018). "Flexible self-organizing maps in kohonen 3.0". In: *Journal of Statistical Software* 87, pp. 1–18. DOI: [10.18637/jss.v087.i07](https://doi.org/10.18637/jss.v087.i07) (cit. on p. 12).
- Weng, S., J. Gou, and Z. Fan (2021). "*h*-DBSCAN: A simple fast DBSCAN algorithm for big data". In: *Asian Conference on Machine Learning*. PMLR, pp. 81–96. DOI: [10.48550/arXiv.2106.09120](https://doi.org/10.48550/arXiv.2106.09120) (cit. on p. 10).
- Xu, R. and D. Wunsch (2005). "Survey of clustering algorithms". In: *IEEE Transactions on Neural Networks* 16.3, pp. 645–678. DOI: [10.1109/TNN.2005.845141](https://doi.org/10.1109/TNN.2005.845141) (cit. on p. 8).
- Zhang, X. et al. (2011). "China Digital Ocean Prototype System". In: *International Journal of Digital Earth* 4.3, pp. 211–222. DOI: [10.1080/17538947.2010.512367](https://doi.org/10.1080/17538947.2010.512367) (cit. on pp. 4, 5).



Masters Program in **Geospatial Technologies**





2000 CLUSTERING OCEAN STRUCTURES USING ARMOR 3D DATA FOR MARINE PATTERN RECOGNITION Miguel Lucas