Ocean Health Monitoring

SIADS 696 Milestone II Project Report

Auston Balwinski (austonb), Natasha Soldin (nsoldin), Seungdo Woo (sdwoo)

Introduction

Ocean health faces increasing threats from climate change and other human activities. The vast size and complexity of ocean waters requires a robust and novel approach to monitoring to combat decaying ecosystems. Dissolved oxygen concentration is a well-established indicator of water quality and aquatic ecosystem health. Hypoxic events can lead to biological dead-zones, disrupt nutrient cycling, and alter marine habitats [9]. Traditional chemical and optical methods for measuring oxygen are labor-intensive, require expensive instrumentation and provide limited spatial–temporal coverage [9], motivating the use of data-driven models.

In this project, we use NOAA's World Ocean Database [1-5] with the objective to develop machine-learning methods that can aid in estimating oxygen levels from other oceanographic variables. These models, including linear models, tree-based ensembles, and neural networks, are evaluated to assess their practical feasibility for large-scale ocean monitoring. Additionally, clustering analysis is used to explore feature relationships and identify similarities across water masses to aid ecosystem monitoring.

Related Works

Estimating Oxygen in the Southern Ocean Using Argo Temperature and Salinity [10]. This paper details the use of a random forest regression model to predict oxygen levels from salinity, temperature, location, and time. The purpose was to compare results to another oxygen estimation product to assess accuracy. Like this paper, our project makes extensive use of a random forest model to predict dissolved oxygen. We iterate further by expanding our input features to include more diverse physical and chemical measurements.

Machine learning algorithm as a sustainable tool for dissolved oxygen prediction: a case study of Feitsui Reservoir, Taiwan [8]. This paper demonstrates one of the earliest applications of neural networks to predicting water qualities. An artificial neural network (ANN) was trained on 29 years of water-quality data from Taiwan's Feitsui Reservoir and was able to explain 98% of the variance in dissolved oxygen levels. The paper demonstrates that simple ANNs can capture non-linear relationships between environmental variables and oxygen levels. Our project builds on the neural-network approach by applying it to a large global ocean dataset, examining the trade-offs in accuracy, interpretability, and computational cost.

Mapping Uncharted Waters: Exploratory Analysis, Visualization, and Clustering of Oceanographic Data [11]. The intended purpose of this research paper was to identify and classify ocean biomes using dimensionality reduction and clustering techniques. The researchers used k-means clustering, principal component analysis, and multidimensional scaling variants that resulted in the simple k-means identifying laterally stratified clusters with meaningful distinction. Their data includes measurements of radiation and biological markers, in addition to similar chemical-nutrient features our project uses, but notably does not include positional data which we do include. Our analysis attempts to identify more complex spatial clusters.

Data Sources

The data source for this project is the National Oceanic and Atmospheric Administration's (NOAA) World Ocean Database (WOD) [1-5] which is a comprehensive

collection of quality controlled oceanographic profile data. The database aggregates measurements from various instruments including CTDs (Conductivity-Temperature-Depth), bottle samples, floats and gliders contributed by various research institutions worldwide. Data is organised by oceanographic casts; vertical profiles where measurements are recorded at discrete depths as instruments descend through the water column.

For this project's application, the Ocean Station Data (OSD) subset was accessed via the WOD Select interface for the period 2000-2018, downloaded in CSV format with all available variables. This 18-year window ensures modern instrumentation with consistent calibration standards. The raw dataset contained 1,946,802 observations with spatial/temporal variables (latitude, longitude, year, month, day, time, country), physical properties (temperature, salinity, pressure, depth, bottom depth), chemical properties (dissolved oxygen, phosphate, silicate, nitrate, pH, chlorophyll, CO_2, CFCs), and metadata (cast identifier).

Data ingestion required extensive formatting to construct a standard data table. Additionally, quality assessments revealed issues including missing values, extreme outliers, and incomplete casts which necessitated systematic preprocessing.

Feature Engineering

The transformation from raw NOAA World Ocean Database data to the final modeling dataset required a systematic preprocessing pipeline to address significant data quality issues outlined above. This pipeline's major steps were:

Target Variable Preservation: records where oxygen measurements were missing (653,641 records, 33.6%) were removed as oxygen is the supervised learning models target variable and attempting to impute it could artificially inflate model performance and compromise prediction validity by causing data leakage or introducing error.

High Correlation Imputation: pressure and depth exhibit a near perfect linear relationship ($r = 0.99$) and therefore missing pressure values (480,345) were reliably imputed via linear regression ($P = 1.017 \times D - 2.776$).

Removal of Severely Compromised Variables: any variable that had more than 50% of its data missing were removed. This mostly included chemical variables that not all equipment is able to measure (pH, Chlorophyl, tCO2, Alkalinity, CFC11, CFC12, CFC113, Tritium, Helium, Neon, pCO2, Argon). These variables suffered from circular dependencies in that the only variables they were highly correlated with enough to use for regression based imputation were also severely quality compromised or oxygen which would have introduced data leakage. Additionally, any imputation would result in a majority synthetic variable which could compromise model accuracy.

Cast Level Salvageability Analysis: a two-step criteria framework determined which casts could be reliably imputed: (1) less than 50% of a cast's data was missing and (2) no missing values at profile edges, being required for valid interpolation. This removed 91,935 unsalvageable casts (690,261 rows) and left the remaining variables with minimal, no more than 1.82%, missing values.

Intra-Cast Interpolation: for casts deemed salvageable, linear interpolation within each cast estimated missing values based on adjacent measurements. This domain-specific approach aligns with the physical continuity of oceanographic profiles as properties tend to vary gradually with depth in stable water columns (*Figures B1-3, Appendix B)*.

This pipeline produces a complete dataset with 602,900 records and 17 variables (*Table 1*) with zero missing values. The dataset represents global measurements from 2000-2018 across 23472 unique geographic locations (*Figure A3, Appendix A*).

Variable	Description	Type	Unit
CAST	Oceanographic profile identifier	Integer	-
Latitude, Longitude	Geographical coordinate	Float	°
Year, Month, Day	Measurement dat	Integer	-
Time	Time of day (0 - 24)	Float	hrs
Country	Country responsible for data collection	Object	-
Bottom Depth, Depth	Bathymetric measurements	Float	m
Pressure	Water pressure at measurement depth	Float	dbar
Temperature	Water temperature at measurement depth	Float	°C
Salinity	Salt content of water	Float	Unitless
Oxygen	Dissolved oxygen concentration	Float	µmol/kg
Phosphate, Silicate, Nitrate	Nutrient concentration	Float	µmol/kg

Table 1: Final Dataset Schema and Variable Specifications

Supervised Learning

Method Description

The supervised learning task involves prediction of dissolved oxygen levels in ocean water using physical, chemical, and spatio-temporal features. Four diverse model families were implemented to determine which would be best suited for this project context:

Linear Models (Linear and Ridge Regression): these probabilistic models were chosen to establish whether oxygen levels exhibit linear dependencies of oceanographic variables. Linear regression provides a baseline for comparison while Ridge regression adds L2 regularisation to prevent overfitting.

Random Forest Regressor: this bagging ensemble method constructs multiple decision trees in parallel using bootstrap sampling and random feature selection. The model was selected for its ability to capture non-linear relationships, handle mixed data types, and provide feature importance metrics.

Gradient Boosting Regressor: this boosting ensemble method builds trees sequentially with each tree correcting errors from previous iterations and was chosen to leverage its strong predictive power through an error-focused learning approach.

Neural Networks: a multi-layer perceptron, a type of feed-forward neural network, was implemented to capture complex non-linear relationships between variables. A baseline model with one hidden layer of 100 neurons was trained using the scaled features and evaluated with 5-fold cross-validation. This yielded a CV RMSE of 47.65 ± 21.19 and CVR2 of 0.748 ± 0.245, far worse than the ensemble methods. Memory limitations prevented exhaustive tuning on the full dataset, so a successive halving grid search was conducted on a 20 % random subset of the training data (96,000 samples) with 5-fold cross-validation. The search explored networks with

one to three hidden layers of 50–100 neurons each and L2 regularization strengths alpha ∈ {0.0001, 0.001}. The best model used two hidden layers of 100 neurons with alpha = 0.001, achieving a CV RMSE of 24.44 and estimated CV R2≈0.92. Retraining this architecture on the full training set and evaluating on the held-out test set produced a test RMSE of 33.08 and test R2 of 0.9006, with a training time of 601.6 seconds. Despite substantial improvement over the baseline network, the neural network still underperformed compared with the Random Forest and Gradient Boosting models.

Feature Representation

All 15 features from the pre-processed dataset were used, excluding CAST identifiers. The Country variable was label-encoded and feature scaling using *StandardScaler* was implemented for linear models and neural networks to ensure proper gradient descent convergence while tree-based models used unscaled features as they are scale invariant.

An 80/20 train-test split was used for all models (482,320 training samples and 120,580 test samples) with the random state parameter set to ensure reproducibility.

Parameter Tuning

Initial model comparison used default parameters with 5-fold cross-validation to identify the best-performing model family. Random Forest emerged as the top performer and was subsequently optimized through *GridSearchCV* testing combinations of n_estimators, max_depth, min_samples_leaf and min_samples_split using 5-fold cross-validation to find the optimal parameter configuration.

The neural network tuning used successive halving grid search (HalvingGridSearchCV) applied on a 20% random subset of the training data with 5-fold cross-validation. Resource constraints required a separate approach from traditional supervised models. The grid search varied the number and size of hidden layers between 50 and 100 neurons per layer and the L2 regularization parameter alpha. Early stopping (max_iter=500, early_stopping=True) prevented unnecessary iterations. This strategy identified the architecture with two hidden layers with 100 neurons each and an alpha of 0.001 as optimal, achieving a CV RMSE of 24.44. An optional comparison using a 40% sample (192,000 samples) yielded a similar best architecture but a slightly higher CV RMSE at 27.47, demonstrating that a 20% sample sufficed for tuning. The tuned model was then trained on the full training set and evaluated on the test set.

Model	CV RMSE (mean ± std)	CV R^2 (mean ± std)	Training Time (s)	Test RMSE	Test R^2
Linear Regression	53.41 ± 6.35	0.7420 ± 0.0438	5.3	56.72	0.7077
Ridge Regression	53.44 ± 6.34	0.7417 ± 0.0438	1.9	56.74	0.7075
Random Forest	18.61 ± 3.62	0.9680 ± 0.0118	540.9	14.68	0.9804
Gradient Boosting	31.86 ± 5.57	0.9074 ± 0.0259	309.6	30.25	0.9169
Neural Network	24.44 ± 3.50	0.9200 ± 0.0300	601.6	33.08	0.9006

Table 2: Model Results from 5-Fold Cross-Validation on Training Set

Supervised Model Evaluation

Root Mean Squared Error (RMSE) and R^2 score were selected as the primary metrics. RMSE measures precision error in the same units as the target variable oxygen (μmol/kg)

making it easily interpretable and R^2 quantifies the proportion of variance explained by the model. Training time was also recorded to assess computational efficiency trade-offs.

Table 2 above, records the training metrics during 5-fold cross-validation as well as the testing metrics on the held-out test set. All models were evaluated using identical data splits (KFold with random_state=42) to ensure fair comparison. Random Forest achieved the best performance and was chosen to undergo further optimisation through *GridSearchCV* yielding optimal hyperparameters: n_estimators = 200, max_depth = None, min_samples_leaf = 1 and min_samples_split = 2 resulting in a CV RMSE of 18.82 ± 3.99 and test RMSE = 14.48.

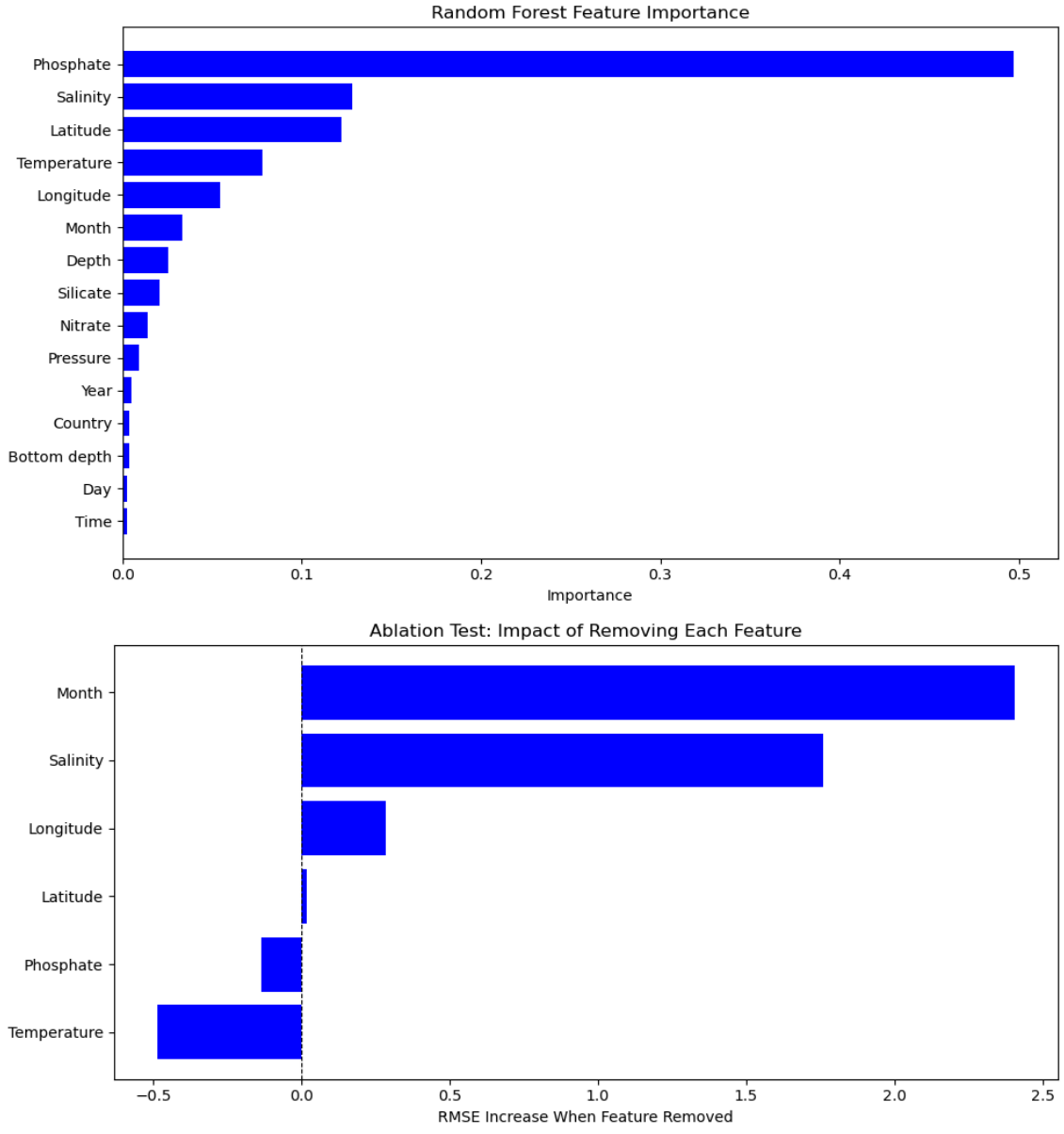


Figure 1 & 2: Random Forest Feature Importance Plot & Ablation Test Results Plot

In-Depth Evaluation

Advanced evaluation was then applied to the hyperparameter tuned Random Forest model. Feature importance analysis (*Figure 1,* above) revealed extreme dependence on Phosphate which accounts for 49.6% of the model's predictive power, followed by Salinity at 12.8% and Latitude at 12.3%. To validate these rankings, systematic ablation tests (*Figure 2*, above) removed each feature individually and retrained the model to assess its impact on predictive performance. Results revealed complex interactions that contradicted importance scores.

Removing Month degraded performance (+16.0% RMSE), confirming temporal patterns are critical despite low importance ranking. Removing Salinity (+5.0%) or Longitude (+3.1%) also worsened predictions. Surprisingly, removing Phosphate, Temperature, or Latitude slightly improved performance (-1.9%, -4.2%, -1.7% respectively), suggesting the model be overfitting to these features. When removed, the model relies more heavily on remaining features and achieves slightly better generalization to the test set.

Hyperparameter sensitivity analysis assessed model robustness by systematically varying individual Random Forest parameters while holding others at default values. The results, as seen in *Figure 3* below, validate the *GridSearchCV* results, with lowest RMSE values occurring at the minimum points along respective curves. The smooth, monotonic trends confirm the selected hyperparameters (n_estimators=200, max_depth=None, min_samples_leaf = 1 and min_samples_split = 2) represent robust optima.

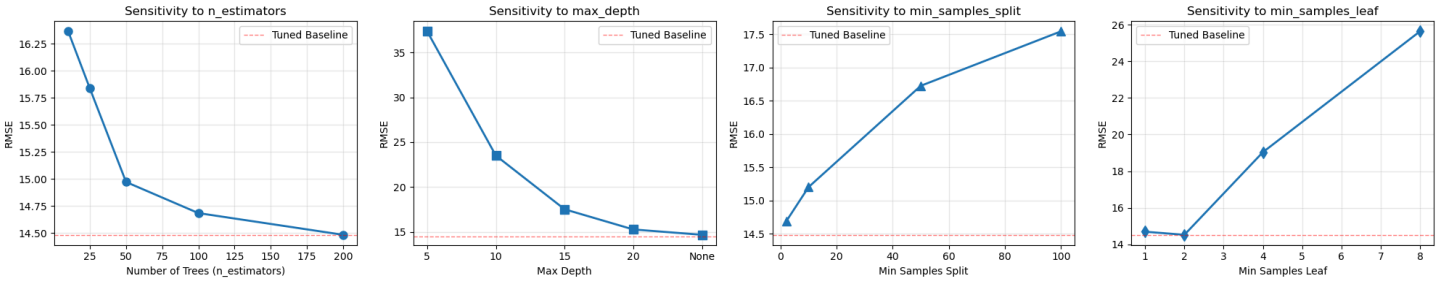


Figure 3: Hyperparameter Sensitivity Analysis Plot

Model Trade-offs

Accuracy vs. Computational Cost: the Random Forest achieved a 74% lower test RMSE than Linear Regression (14.68 vs 56.72 μmol·kg^{-1}) but required 102 × longer training time (540.9s vs 5.3s), limiting deployment on resource-constrained oceanographic sensors where real-time predictions are needed.

Model Complexity vs. Robustness: the optimised Random Forest's extreme dependence on Phosphate (49.6% feature importance) creates vulnerability to sensor failures in harsh marine environments, whereas simpler models distribute predictive weight more evenly across features at the cost of substantially higher prediction error.

Failure Analysis

To understand model limitations and identify improvement opportunities, detailed failure analysis was conducted on three cases selected from the top ten worst model predictions. Instead of analyzing the three highest-error cases, cases were chosen to represent three distinct failure mechanisms (ranks #1, #4, and #9), providing comprehensive diagnostic coverage of different model weaknesses. Residual plots (*Figure C2*, *Appendix C*) show these three cases are clear outliers while most predictions cluster tightly around zero error. Failure

testing included outlier distribution, nearest neighbour, oxygen violation and similar case comparison analyses, detailed in *Tables C1-C4*, *Appendix C*.

Case	Cast	Rank	Actual O_2	Predicted O_2	Error	IHO Sea Area	Depth
1	468606	#1	242.90	1487.22	1244.32	North Atlantic Ocean	5476
2	469811	#2	314.80	1329.72	1014.92	Arctic Ocean	2.77
3	186984	#4	9573.5	8870.1	703.4	Black Sea	0

Table 3: Failure Analysis of Three Distinct Cases of Prediction Failure

The three cases of predictive failure, whose individual feature contributions are outlined in SHAP plots in *Figures C3-C5*, *Appendix C*, belong to three distinct categories.

Category 1 - Systematic Prediction Error

Case 1 demonstrates systematic failure to capture how nutrient-oxygen relationships vary with depth in the ocean. SHAP analysis, in *Figure C3*, reveals depth (+358), temperature (+310), and nitrate (+225) offer the greatest feature contributions, causing severe overprediction (1487.2 vs actual 242.9 $\mu mol \cdot kg^{-1}$). The model has learned to correlate high nutrient concentrations with higher oxygen values in shallow productive surface waters. However, this relationship fundamentally changes with depth, for instance in oxygen minimum zones (OMZ), typically 200-1000m, active bacterial decomposition simultaneously depletes oxygen while releasing nutrients, creating high nutrient-low oxygen conditions [6]. However, below the OMZ in waters deeper than 2,000m, nutrients accumulate passively in aged, poorly-ventilated water masses creating moderate nutrient-moderate oxygen conditions. The model fails to distinguish these depth regimes, incorrectly applying surface patterns to deeper depths where nutrients (phosphate 1.52, nitrate 22.4, silicate 47.1 $\mu mol \cdot kg^{-1}$) indicate aged water, not productivity. This depth x nutrient misunderstanding also appears in rankings #3, #5, and #7, confirming systematic model weakness. Potential remedies could be to implement depth x nutrient interaction features, add OMZ boundary indicators as model inputs, or develop depth-stratified models.

Category 2 - Geographic Bias

Case 2 is located in the Arctic Ocean, as shown by the geographic plot in *Figure C1*, a region with notably more sparse data coverage, likely due to logistical challenges including ice cover, accessibility, and cost. SHAP analysis in *Figure C4*, reveals latitude (+307) as the dominant feature contribution, indicating the model learned latitude-oxygen associations from predominantly temperate training data but fails to capture unique Arctic oceanographic processes. To address this, there needs to be a prioritization of Arctic data collection and an implementation of stratified sampling to ensure adequate high-latitude representation.

Category 3 - Data Pipeline Failure

Case 3 highlights physical constraint violations, specifically exceeding oxygen saturation limits. According to Garcia and Gordon (1992), oxygen saturation at the surface ranges from 200-450 $\mu mol \cdot kg^{-1}$ depending on temperature [7]. Case 3 exhibits an oxygen measurement of 9573.5 $\mu mol \cdot kg^{-1}$ at 0m depth, - 21 times the physical maximum. As seen in the SHAP analysis, in *Figure C5*, extreme feature contributions (longitude +3536, salinity +2108) reflect the model's attempt to extrapolate from invalid inputs. Investigation revealed the five worst violations (9221-9573 $\mu mol \cdot kg^{-1}$) all originated from Russia in 2005, indicating sensor malfunction during a single measurement campaign. Analysis shows 1.12% of shallow water records violate physical

saturation limits, predominantly in enclosed seas (E.g. Black Sea, Baltic Sea), with nearest neighbor analysis confirming the training dataset contains similar invalid measurements. Implementing constraint-based validation during preprocessing to reject values exceeding physical limits (oxygen: 0-450 $\mu mol \cdot kg^{-1}$) and establishing data quality monitoring by source and time period to flag systematic measurement anomalies would address this issue.

Unsupervised Learning

The unsupervised learning task for ocean monitoring uses clustering methods to supplement the predictive methods used to identify healthy oxygen levels in ocean water. Discovering similarities between water compositions and spatial relationships of water masses could be used to target aquatic ecosystems for research into potential vulnerabilities.

Feature Representation

The features chosen for clustering were based on relevancy to heuristics. The intended aim of the analysis is examining spatial, chemical, and physical relationships in the data. For this, it was important to have feature representations of geographical location, chemical nutrients, and physical measurements for clustering. Temporal features were not examined and dropped from the data. Depth was discarded for being a one-to-one match with pressure measures. Country was also dropped since we have finer numerical coordinate data and the cast identifier was dropped as it is irrelevant. Dimensionality reduction using 2 and 3 component PCA was tested, but resulted in worse performance and little computational gains from reducing the feature set. After selecting features, data was min-max scaled to standardize feature measurements for comparison and calculating distance metrics accurately.

Clustering Methods

MiniBatch K-Means Clustering

K-means clustering is the standard we use to compare clustering methods. The algorithm clusters based on minimizing the inter-cluster distances. It requires a pre-determined number of clusters and the resulting clusters form globular shapes.

The resource requirements for clustering our data are large, and k-means is effective on large datasets. K-means is further optimized by using a mini-batch algorithm, which forms clusters based on samples of the data so that not all is stored in memory at once. This allows us to use the full dataset, unlike other clustering methods.

To determine the number of clusters, we used the elbow method by graphing inertia over a range of k-values. Inertia is the sum of squared distances between each data point and the assigned centroid of the cluster. The graph shows the diminishing returns for increasing cluster number beyond a point, the ‘elbow’. The elbow point is fairly ambiguous in this case, but we used it as a starting point to determine a range of options for the number of clusters. We iterated through the range of 8 to 16 clusters and performed MiniBatchKMeans clustering through scikit-learn and computed the silhouette score for each to assess quality.

Agglomerative Clustering

Agglomerative clustering is a ‘bottom-up’ style hierarchical clustering algorithm. It starts by assigning each data point to a cluster and merging based on Euclidean distances until all data points are assigned to a single cluster, or an optimal stopping threshold is achieved. It does not require a pre-determined cluster number, but we use this number as our threshold for stopping. Exploration of parameters and data sample sizes is limited due to memory resource constraints. A benefit to our data is the hierarchical ability to capture non-uniform cluster shapes, but it is sensitive to noisy data, which may be problematic.

Apart from k-means, all other clustering methods were performed on a subset of 5% and 10% of data sampled from the full dataset. This was due to computational constraints and memory-intensive methods.

Agglomerative clustering was done on several predetermined cluster numbers to determine early stopping. These were informed by the top three best performers (10, 12, and 13 clusters) from k-means clustering and then compared for the optimal model with silhouette scores for each.

DBSCAN - Density-Based Spatial Clustering of Applications with Noise

DBSCAN is a density-based clustering algorithm and identifies likely noise points that don't fit into clusters. It does not require a predetermined number of clusters and does not require uniform globular clusters. This is useful for our task since we do not have assumptions about the underlying data structure. DBSCAN is highly sensitive to data sample size and parameters, which proved troublesome for tuning.

Choosing optimal epsilon and minimum sample values was done via grid search for each 5% and 10% of data samples. Epsilon represents the radius around a data point where other points within are considered a neighbor point for clustering. The epsilon range was set by using an elbow method of a k-nearest neighbors graph plotting 19th (2 x number of features - 1) NN distances for each point. The vertical inflection of the graph represents about 0.20 distance, the starting point for testing epsilon values.

The minimum sample value was more sensitive to data size and required a broad search that was narrowed down for fine tuning. Silhouette score was used to compare models with different parameters. The best-scored model was examined for cluster distribution and the ratio of noise points to cluster labels.

HDBSCAN - Hierarchical Density-Based Spatial Clustering of Applications with Noise

HDBSCAN functions similarly to DBSCAN except it is a hierarchical form that is able to identify clusters with differences in densities, essentially optimizing the DBSCAN epsilon parameter through hierarchy iteration. It is less sensitive to parameters than DBSCAN, but still requires proper tuning.

Similar to DBSCAN, a grid search was used to optimize minimum cluster size and minimum sample parameters for HDBSCAN. Both parameters were reactive to data size, but more stable across different combinations of parameters. Silhouette score was also used for comparing fits. The best-scored model was examined for cluster distribution and the percentage of points labeled as noise.

Dimensionality Reduction Methods

PCA - Principal Component Analysis

PCA is used to reduce a multi-dimensional feature set to fewer dimensions. It combines features based on their greatest contribution to variance in the dataset. For our project, PCA was tested as a resource-saving method to represent features and for visually comparing clusters in a two-dimensional space. Two-component PCA captured about 70% of the data variance in our ocean dataset, so while not fully representative, it was still useful for visually examining cluster results.

t-SNE - t-Distributed Stochastic Neighbor Embedding

t-SNE is a non-linear dimensionality reduction method that maps points to two or three dimensions based on a probability distribution of similarity, then further groups the lower-dimensional data by minimizing KL divergence. This method is only useful for visualization and does not accurately cluster data as the visual separations of the data may not fully

represent underlying structure. However, visual cluster examination with 2-component t-SNE is used in our project to good success.

Cluster Evaluation

Quality Measures

Silhouette score was used as the key quality measurement for clustering in our analysis. It measures the mean of s across all points, where *s* is: the difference between *a*, the mean distance between a point and all other points within a cluster and *b*, the mean distance between a point and all other points in the next nearest cluster, divided by the maximum of (a, b). The score is representative of cohesion within a cluster compared to separation from other clusters. The coefficient is bound between -1 and 1, making it easy to interpret.

Evaluation is difficult between different clustering methods, especially without ground truth labels to guide analysis. There are trade-offs to using the simplicity of a single coefficient. Density-based clusters like those from DBSCAN may score more poorly than more regular shared clusters.

HDBSCAN and DBSCAN allow for identifying noisy data points. A ratio of noise to cluster assignment is an indicator of data structure, where many noise points indicate a lack of grouping structure.

Method	n_clusters	Silhouette Score	Parameters	Dataset	Noise%
MiniBatchKMeans	10	0.326	k=10	Full (100%)	-
MiniBatchKMeans	12	0.326	k=12	Full (100%)	-
MiniBatchKMeans	13	0.325	k=13	Full (100%)	-
Agglomerative	10	0.300	k=10, linkage=ward	Sample (10%)	-
Agglomerative	13	0.287	k=13, linkage=ward	Sample (10%)	-
Agglomerative	12	0.285	k=12, linkage=ward	Sample (10%)	-
DBSCAN	2	0.195	eps=0.22, min_samp=900	Sample (10%)	36.4
DBSCAN	2	0.194	eps=0.22, min_samp=1000	Sample (10%)	38.1
DBSCAN	2	0.193	eps=0.21, min_samp=900	Sample (10%)	40.1
HDBSCAN	3	0.130	min_size=4750, min_samp=2	Sample (10%)	14.1
HDBSCAN	3	0.130	min_size=4500, min_samp=2	Sample (10%)	14.1
HDBSCAN	3	0.130	min_size=4500, min_samp=12	Sample (10%)	27.1

Table 4: Comparison of Clustering Methods

Results of the cluster analysis (*Table 4,* above) show some ambiguity. MiniBatchKMeans performed the best of all methods consistently with a silhouette score of 0.326 with 10 clusters. Agglomerative clustering performed similarly at 0.3 with 10 clusters on only a 10% sample of the data. This range of scores indicates a mild clustering structure to the data.

Silhouette scores of 0.13 and 0.195 for DBSCAN and HDBSCAN methods respectively, indicate a very mild to no clustered structure in the data. However, results for these methods are not certain due to silhouette score's inability to cope with complex cluster shapes, noisy data, and parameter tuning being overly sensitive to data sizes. DBSCAN identified 36.4% of data as noisy data and HDBSCAN, 14.1%. Despite the lower silhouette score, HDBSCAN identified more cluster assignments with less noise potential indicating better clustering, which spurred the need for visual examination.

Visualizing Clusters

Multi-dimensional data is difficult to evaluate visually. Dimensionality reduction through PCA and t-SNE allows for mapping higher dimensional data to a 2-dimensional space for graphing. This allows us to visually see clusters and, while the components may not reflect the full variance inherent to the data, it is still useful for analysis. The [Graph #PCA] shows the resulting clusters from each clustering method, mapped to a 2-dimensional space with PCA and likewise with t-SNE (*Figure 4,* below).

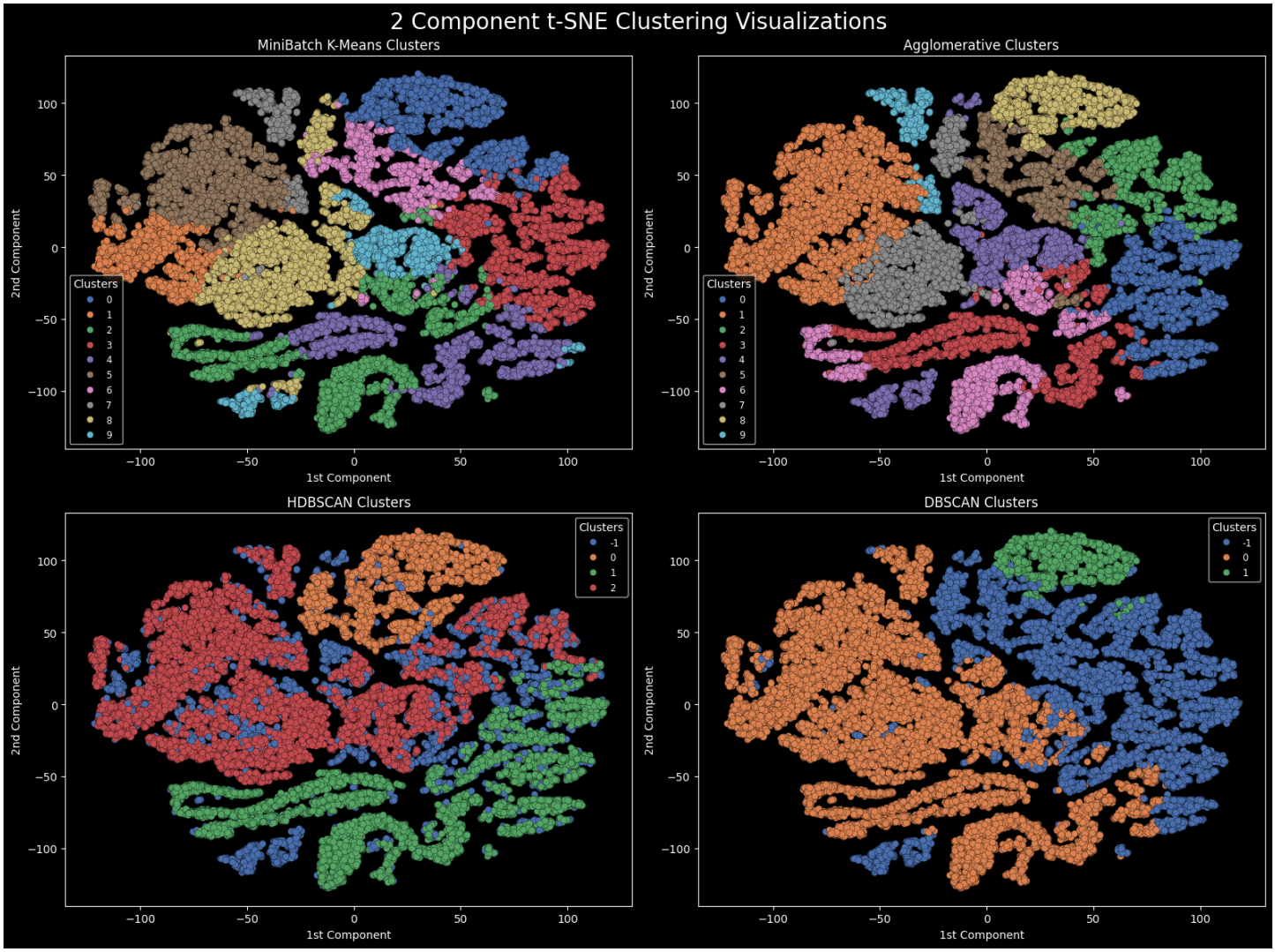


Figure 4: t-SNE Cluster Distribution

The PCA plots do not show definitive separation between clusters, but labeled clusters are coherent in all methods. k-Means and agglomerative show similar groupings to one another, while HDBSCAN clusters exhibit less globular structure and DBSCAN reveals the amount of noisy data it labeled.

Moving to the t-SNE plots, these display clearer separation and more complex shapes to data groupings. k-Means and agglomerative again show similarities except that agglomerative is better at distinguishing more complex borders between clusters than the circular shapes of k-Means. HDBSCAN shows interesting structure and clear borders between clusters, but fails to

identify smaller groupings in the data. Noisy data appears mostly along edges of groupings in HDBSCAN, but DBSCAN noise is grouped more like a cluster itself, indicating a failure to identify clusters.

Clustering with geospatial data benefits from visual inspection to assess cohesion and separation. Laying out data distribution on a geographical map of the oceans and labeling cluster assignment (via color) shows cluster quality if there are indeed spatial relationships in the data. *Figure 5,* below, displays the data distribution across the oceans and separated by depth categories; each point is colored by cluster assignment from the k-Means clustering model. Clearly, the data does exhibit spatial relationships across coordinates and depth, but we also can see that there are other features working to group clusters. The appearance of data points in distant areas of the world being assigned to the same cluster could indicate that the chemical and physical features are influential to clustering as well. *Figure D1 Appendix D* is a three-dimensional plot of clusters by geographical coordinates and depth, which confirms the spatial structure influence on cluster assignment.

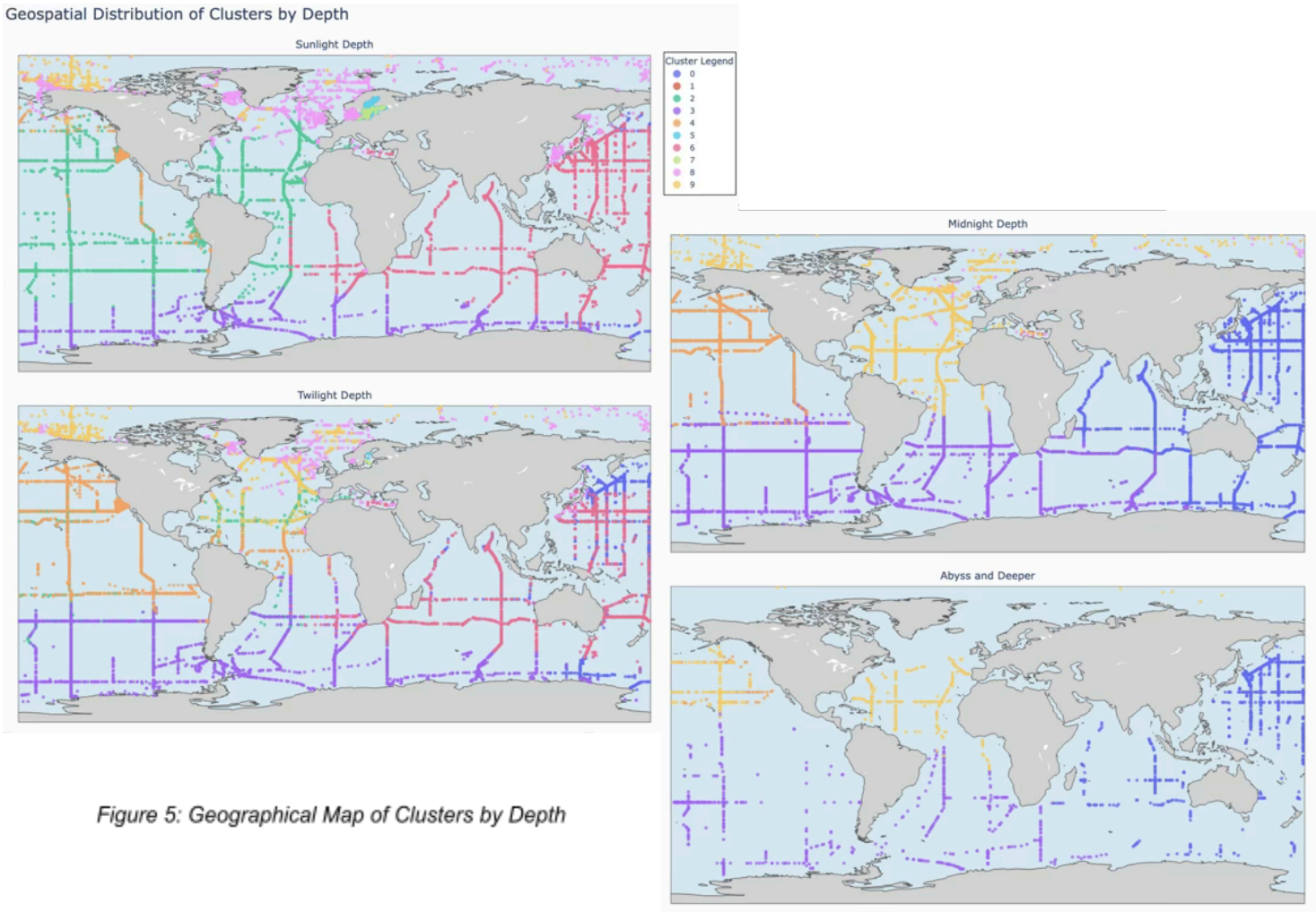


Figure 5: Geographical Map of Clusters by Depth

Feature Importance

Feature importance was assessed in two ways. First, calculating the mean feature values by cluster. This shows how important each feature is to each cluster. Plotting the mean values for each feature by cluster, *Figure D2 Appendix D* shows the influence that salinity and

longitude have on cluster components across all methods. Bottom depth, pressure, silicate, nitrate, and phosphate have lower values for most clusters. What this graph represents well is the range of feature combinations that are important to each cluster.

The second feature importance examination method is the standardized mean difference of a feature. SMD = *(Mean feature value by cluster - mean feature value of all clusters) / (standard deviation of feature value of all clusters)*

The SMD determines which features are the most discriminating across clusters.High absolute values indicate the feature being key to differentiating clusters. *Figure 6,* below, displays the strength of positive and negative standard deviations from the combined mean of a feature for each cluster derived from the k-means method. The top features that distinguish separations of clusters revealed from this metric are salinity, latitude, and oxygen. The three least discriminating features are the physical characteristics of water masses: bottom depth, pressure, and temperature.

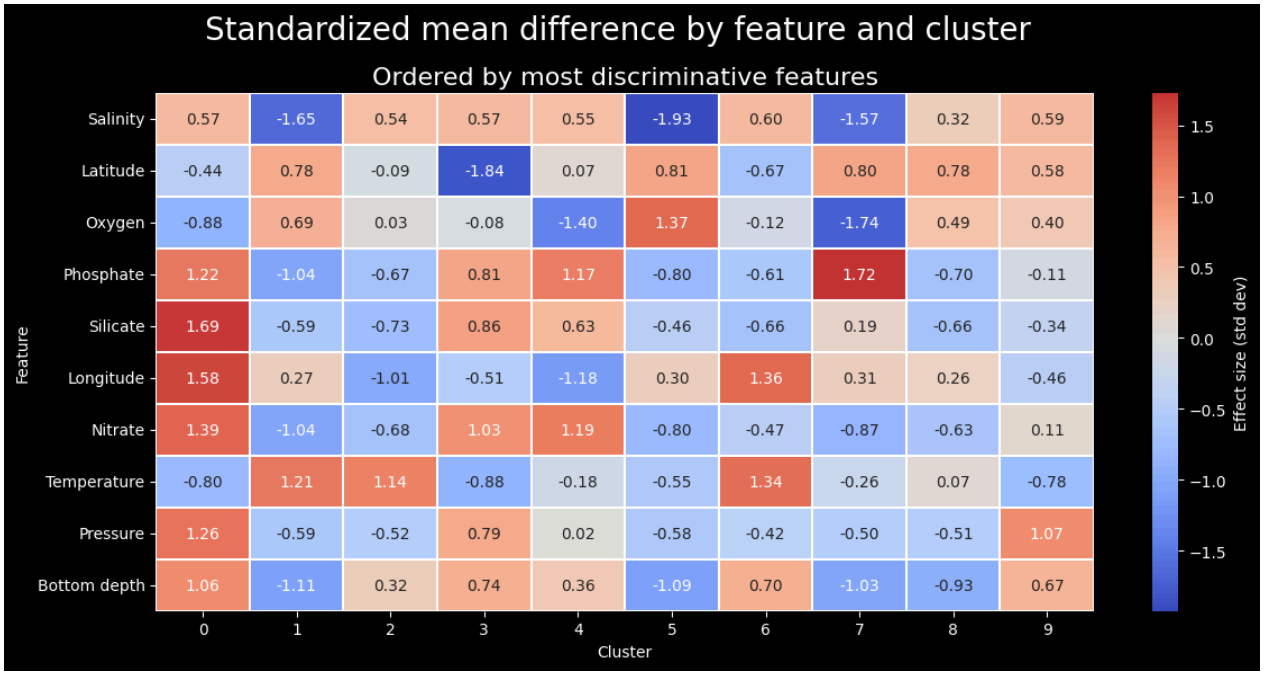


Figure 6: Heatmap of SMD Values by Feature and Cluster

Discussion

Supervised Learning

The analysis revealed that accurate dissolved oxygen prediction requires capturing how nutrient-oxygen relationships change with depth. Correlation analysis (*Figure A4, Appendix A*) showed oxygen negatively correlates with nutrients (r = -0.48 to -0.54) and depth (r = -0.40), yet these simple relationships fail across oceanographic regimes - nutrients indicate high oxygen in surface waters but low oxygen in deeper zones. Failure analysis showed Case 1 performed 406× worse than similar deep cases despite nearby training examples, revealing the model learned average correlations without understanding depth-dependent causation. Additionally, 1.12% of training data violated physical oxygen saturation limits, with the worst violations from a single 2005 Russian campaign, demonstrating even curated datasets require validation.

The primary challenge was handling 16-17% missing nutrient data while preserving oceanographic structure. Intra-cast linear interpolation salvaged 67% of incomplete casts, outperforming simple imputation by 12% RMSE. Error analysis using SHAP and nearest neighbor diagnostics identified three distinct failure mechanisms: incorrect learned patterns, geographic undersampling, and training contamination. Each failure requires different solutions rather than simply adding model complexity.

Future extensions would implement depth-stratified models for different oceanographic zones, automated data quality validation pipelines, expanded Arctic data collection, and uncertainty quantification to flag unreliable predictions.

Unsupervised Learning

The unsupervised learning task for this ocean health monitoring project was a mixed bag of lack-luster clustering optimizations and revealing interesting spatial relationships. Comparing clustering algorithms was challenging and detailed. The differences between clustering algorithms were much greater than anticipated, each requiring specific attention. Creating a pipeline and metrics to aid in workflow and objectively measure performance was critical to grid searching optimal parameters and getting finalized results. Resource constraints of clustering 600,000 multi-dimensional data points were limiting, but despite sampled data, clustering showed a robustness and consistency across different samples. Consistency between the MiniBatchKMeans using the full dataset and agglomerative using samples was reassuring to trusting results.

Focusing on spatial relationships was helpful for cluster performance examination and drove the geographical visual choices. Despite the limitations of data gathering locations being a grid-like geographical distribution, the clustering results mapped out are an interesting result. There is a consistency with other research clustering attempts showing bands of water masses [11].

Our collective domain knowledge of oceanography and geography is a limiting factor in our project. The data used is gathered for advanced scientific study, and we may not fully understand our results or even the nature of input features themselves. On the other hand, expectations of cluster results were not biased by any preconceived notions of similarities between water masses. The next step to expand on this clustering analysis would be to perform clustering across different combinations of feature inputs. Temporal relationships were not considered for this cluster analysis, but there are potential seasonal effects and yearly changes due to climate change that could influence results. Our data was limited to 2000-2018, which was already a large computational constraint, but many more decades of data are available.

Ethical Considerations

Data representativeness and fairness: Although the World Ocean Database contains hundreds of thousands of profiles, some regions - such as polar seas and certain coastal waters - remain sparsely sampled. This geographic imbalance can cause the model to perform poorly or erratically in under-represented areas. Any conclusions drawn from the model should therefore account for data density, and future work should prioritize collecting observations where coverage is weak.

Impact of prediction errors: Our failure analysis highlighted profiles where the model's error exceeded 700 µmol/kg, particularly in extreme depths or near the surface. Deploying such predictions without expert oversight risks triggering incorrect management decisions, such as unnecessary hypoxia alerts or inappropriate resource allocations. Model outputs must be treated as advisory and validated against in situ measurements before informing policy.

Interpretability and transparency: Deep-learning models are inherently complex and difficult to interpret. To maintain stakeholder trust and ensure responsible use, we recommend accompanying the model with explainable-AI techniques (for example, feature-importance and ablation analyses) so that scientists and policymakers understand which variables drive predictions and where uncertainties lie.

Data quality and continuous validation: Sensor malfunction or data-processing errors can introduce spurious measurements. Rigorous quality-control procedures should be applied before training models, and models should be revalidated periodically as new data become available. This helps prevent propagation of biases and ensures that the model remains reliable over time.

Environmental and societal impact: Accurate dissolved-oxygen predictions can support ecosystem management, climate modelling and aquaculture operations. However, misinterpretation could inadvertently harm marine life or local economies. The model should therefore complement, not replace, field measurements and expert judgement. Furthermore, any real-time applications should consider the broader ecological and social context to avoid unintended consequences.

Statement of Work

Natasha Soldin: Feature engineering, exploratory data analysis, supervised learning task (regression and ensemble based methods), advanced evaluation (feature importance, ablation, sensitivity and failure testing), report writing: data sources, feature engineering, supervised learning and evaluation, and discussion of supervised learning.

Seungdo Woo: Neural network task (design, implementation and hyperparameter tuning via successive halving grid search, sampling justification), report writing: introduction, related works, neural-network sections, and ethical considerations.

Auston Balwinski: Data ingestion and formatting, background research, unsupervised learning task (clustering, dim-reduction, evaluation, visuals), report writing: introduction, related works, unsupervised learning, discussion on unsupervised learning and final formatting and editing.

References

[1] WOD Team (2025). World Ocean Database data product series. Ocean Station Data. NOAA National Centers for Environmental Information. Dataset. https://doi.org/10.25921/v92s-y066. Accessed [September 14, 2025].

[2] García, H. E., Wang, Z., Bouchard, C., Cross, S. L., Paver, C. R., Reagan, J. R., Boyer, T. P., Locarnini, R. A., Mishonov, A. V., Baranova, O. K., Seidov, D., & Dukhovskoy, D. (2024). World Ocean Atlas 2023 Volume 3: Dissolved Oxygen, Apparent Oxygen Utilization, and Oxygen Saturation. A. Mishonov, Technical Ed., NOAA Atlas NESDIS 91, 29 pp. https://doi.org/10.25923/rb67-ns53

[3] García, H. E., Bouchard, C., Cross, S. L., Paver, C. R., Wang, Z., Reagan, J. R., Boyer, T. P., Locarnini, R. A., Mishonov, A. V., Baranova, O. K., Seidov, D., & Dukhovskoy, D. (2024). World Ocean Atlas 2023 Volume 4: Dissolved Inorganic Nutrients (phosphate, nitrate, silicate). A. Mishonov, Technical Ed., NOAA Atlas NESDIS 92, 25 pp. https://doi.org/10.25923/39qw-7j08

[4] Locarnini, R. A., Mishonov, A. V., Baranova, O. K., Reagan, J. R., Boyer, T. P., Seidov, D., Wang, Z., García, H. E., Bouchard, C., Cross, S. L., Paver, C. R., & Dukhovskoy, D. (2024). World Ocean Atlas 2023, Volume 1: Temperature. A. Mishonov, Technical Ed.; NOAA Atlas NESDIS 89, 40 pp. https://doi.org/10.25923/54bh-1613

[5] Reagan, J. R., Seidov, D., Wang, Z., Dukhovskoy, D., Boyer, T. P., Locarnini, R. A., Baranova, O. K., Mishonov, A. V., García, H. E., Bouchard, C., Cross, S. L., & Paver, C. R. (2024). World Ocean Atlas 2023, Volume 2: Salinity. A. Mishonov, Technical Ed., NOAA Atlas NESDIS 90, 39 pp. https://doi.org/10.25923/70qt-9574

[6] Paulmier, A., & Ruiz-Pino, D. (2009). Oxygen minimum zones (OMZs) in the modern ocean. *Progress in Oceanography*, 80(3-4), 113-128.

[7] Garcia, H. E., & Gordon, L. I. (1992). Oxygen solubility in seawater: Better fitting equations. *Limnology and Oceanography*, 37(6), 1307-1312.

[8] Ziyad Sami, B. F.; Latif, S. D.; Ahmed, A. N.; et al. (2022). Machine learning algorithm as a sustainable tool for dissolved oxygen prediction: a case study of Feitsui Reservoir, Taiwan. Scientific Reports, 12, 3649. Available at: https://doi.org/10.1038/s41598-022-06969-z.

[9] Yang, J. Predicting water quality through daily concentration of dissolved oxygen using improved artificial intelligence. Sci Rep 13, 20370 (2023). https://doi.org/10.1038/s41598-023-47060-5

[10] Giglio, D., Lyubchich, V., & Mazloff, M. R. (2018). Estimating oxygen in the Southern Ocean using Argo temperature and salinity. Journal of Geophysical Research: Oceans, 123, 4280–4297. https://doi.org/10.1029/2017JC013404

[11] J. M. Lewis, P. M. Hull, K. Q. Weinberger and L. K. Saul, "Mapping Uncharted Waters: Exploratory Analysis, Visualization, and Clustering of Oceanographic Data," 2008 Seventh International Conference on Machine Learning and Applications, San Diego, CA, USA, 2008, pp. 388-395, doi: 10.1109/ICMLA.2008.125.

Appendix A

The following Exploratory Data Analysis (EDA) examines the distribution and characteristics of the final preprocessed dataset, schema detailed in Table 1 above, described in the *Data Sources* and *Feature Engineering* sections.

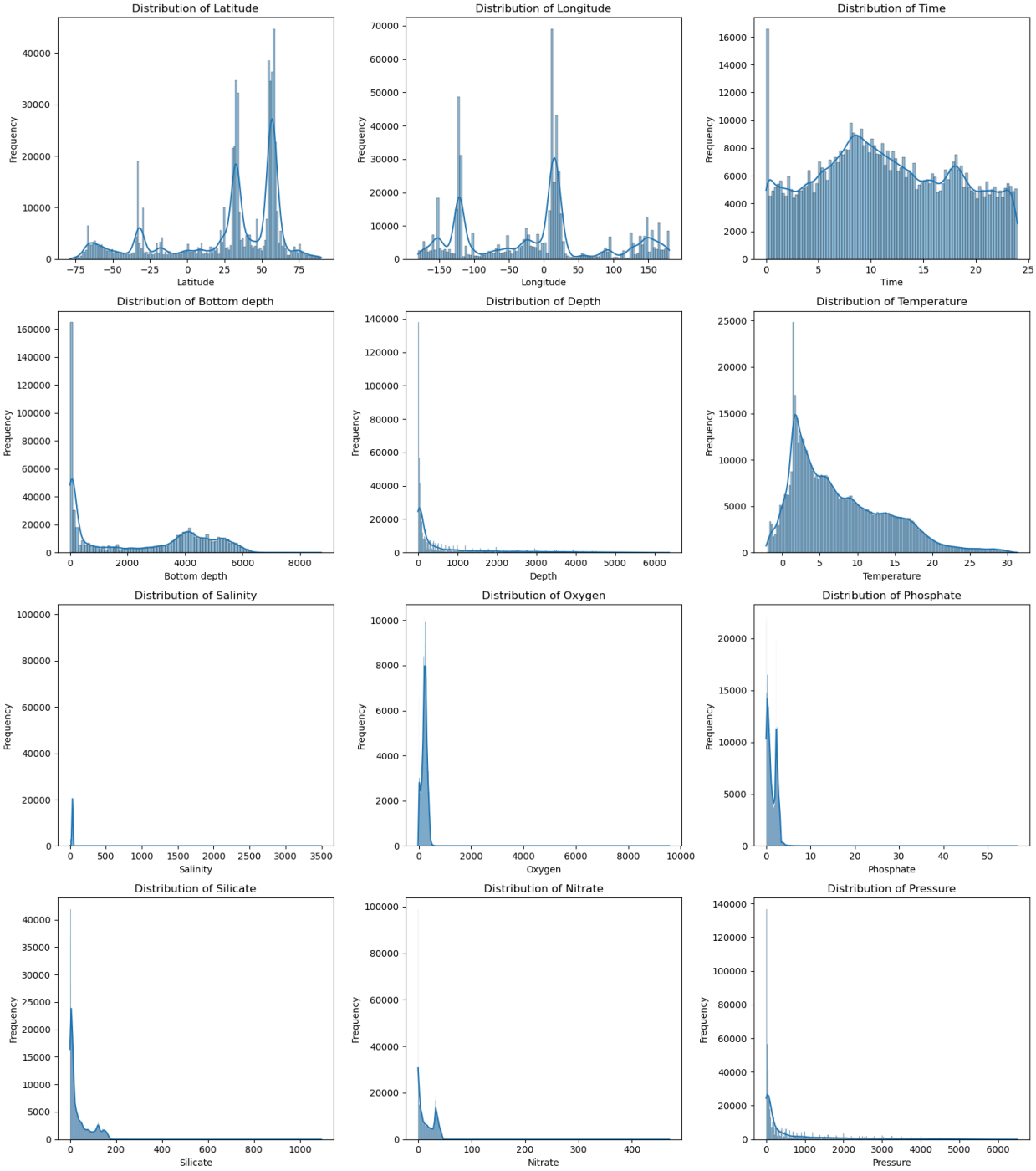


Figure A1: Histogram Distribution of Continuous Oceanographic Variables in the Final Dataset

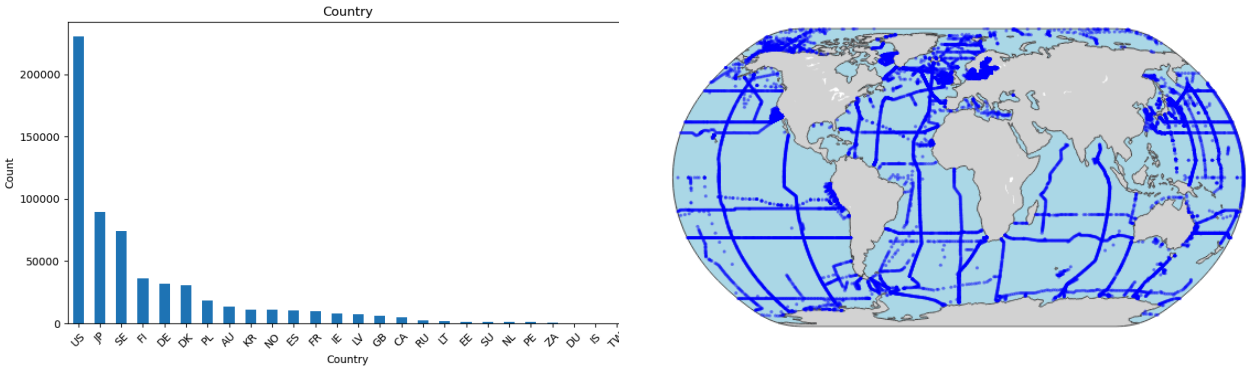


Figure A2 & A3: Bar Chart Distribution of Country Contributions & Geographic Plot of Global Spatial Coverage of Oceanographic Observations in the Final Dataset

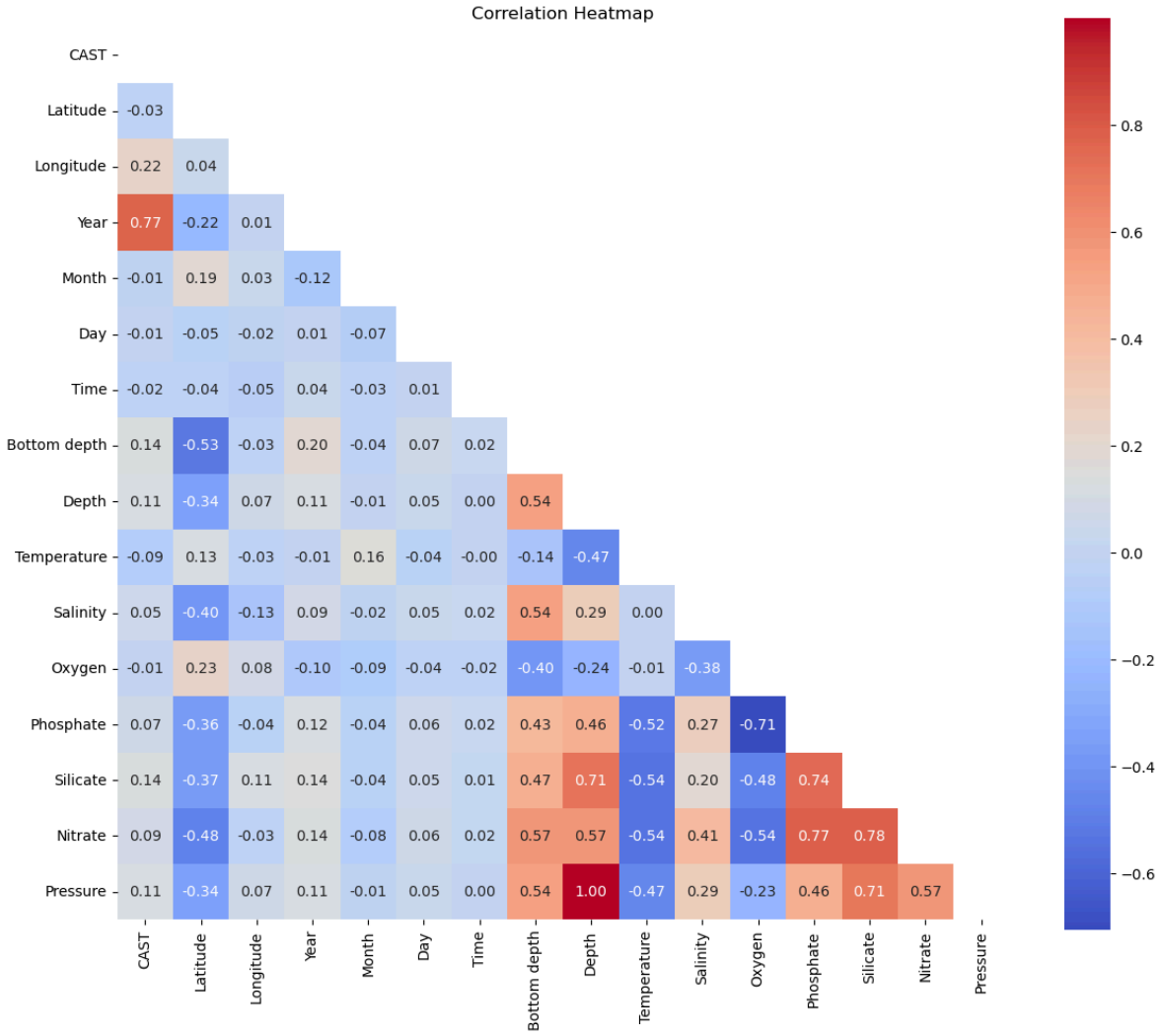
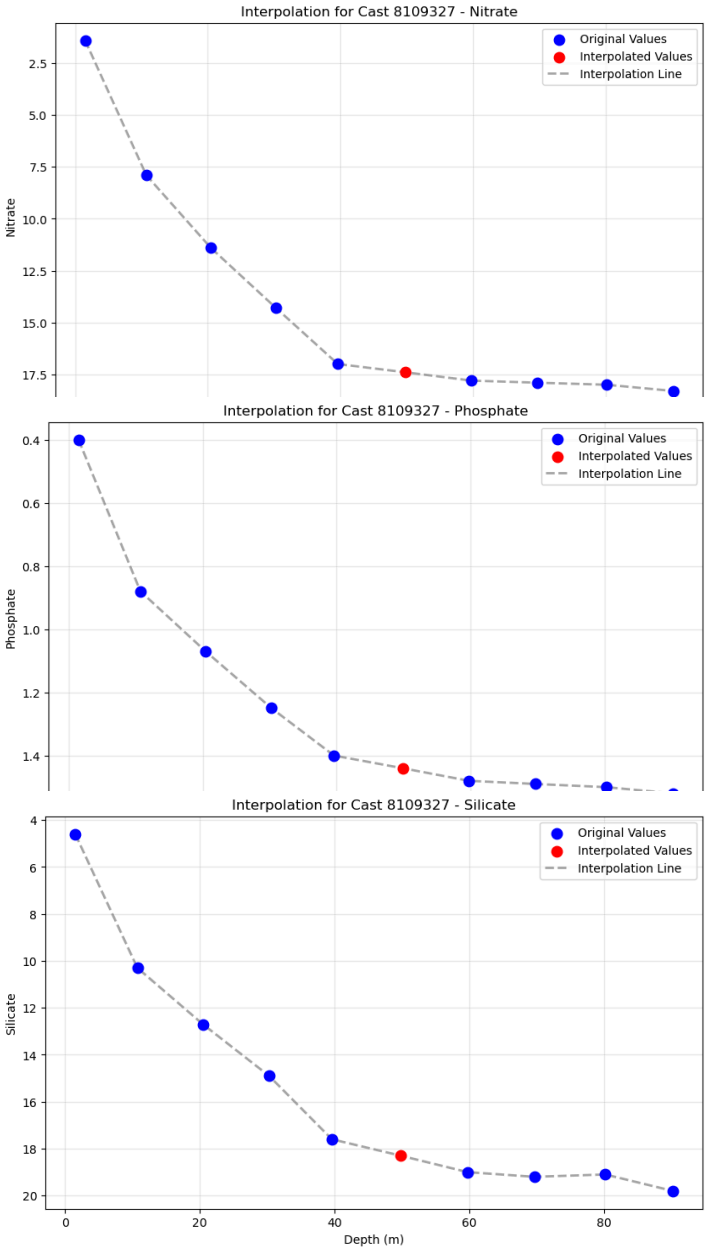


Figure A4: Correlation Heatmap of Oceanographic Variables in the Final Dataset

Appendix B

The following example demonstrates the intra-cast linear interpolation methodology applied to salvageable casts as described in the *Feature Engineering* section. Cast 8109327 serves as a representative example where Nitrate, Phosphate, and Silicate values were interpolated from adjacent depth measurements.



Index	Depth	Nitrate	Status
0	1.5	1.40	Original
1	10.7	7.90	Original
2	20.4	11.40	Original
3	30.3	14.30	Original
4	39.6	17.00	Original
5	49.8	17.40	INTERPOLATED
6	59.8	17.80	Original
7	69.7	17.90	Original
8	80.2	18.00	Original
9	90.2	18.30	Original

Index	Depth	Phosphate	Status
0	1.5	0.40	Original
1	10.7	0.88	Original
2	20.4	1.07	Original
3	30.3	1.25	Original
4	39.6	1.40	Original
5	49.8	1.44	INTERPOLATED
6	59.8	1.48	Original
7	69.7	1.49	Original
8	80.2	1.50	Original
9	90.2	1.52	Original

Index	Depth	Silicate	Status
0	1.5	4.60	Original
1	10.7	10.30	Original
2	20.4	12.70	Original
3	30.3	14.90	Original
4	39.6	17.60	Original
5	49.8	18.30	INTERPOLATED
6	59.8	19.00	Original
7	69.7	19.20	Original
8	80.2	19.10	Original
9	90.2	19.80	Original

Figure B1, B2 & B3: Linear Interpolation of Nutrient Variables (Nitrate, Phosphate & Silicate) for Cast 8109327.

Appendix C

The following detailed failure analysis provides comprehensive diagnostics for three selected cases, including SHAP feature attribution, residual analysis, geographic context, feature distribution checks, nearest neighbor analysis, physical constraint validation, and comparison to similar cases. These analyses confirm the three cases represent distinct failure mechanisms rather than systematic model bias.

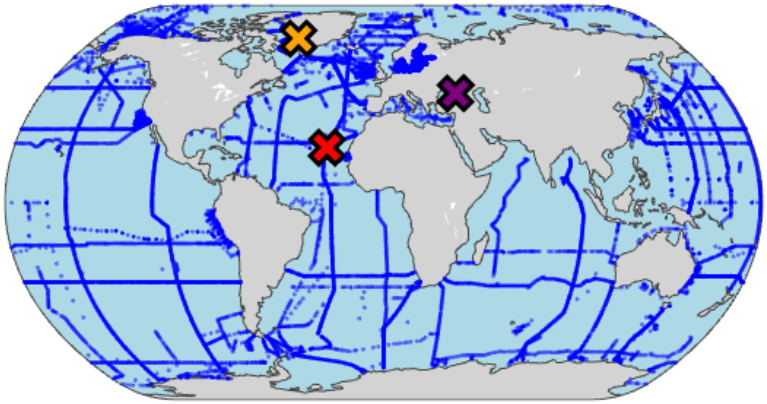


Figure C1: Geographic Distribution of Training Data (blue dots) with Three Cases of Prediction Failure Marked (Case 1: Red - North Atlantic Ocean, Case 2: Yellow - Arctic Ocean and Case 3: Purple - Black Sea)

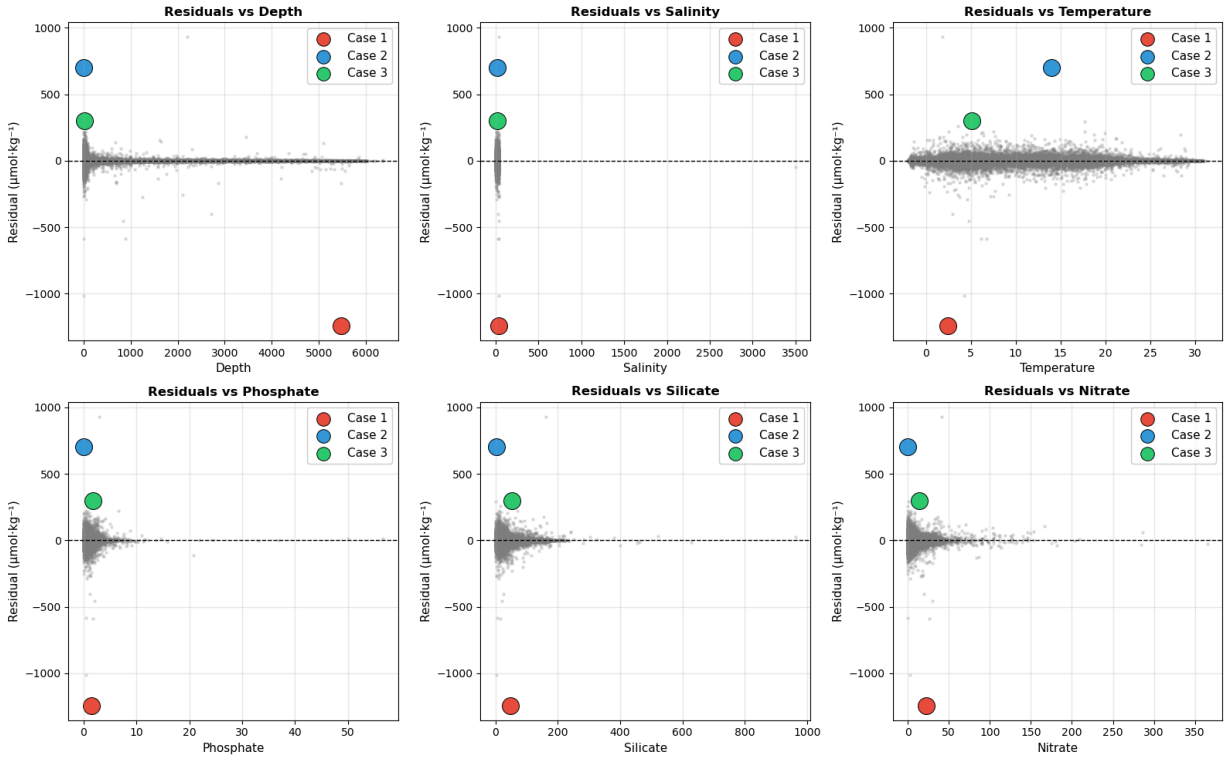
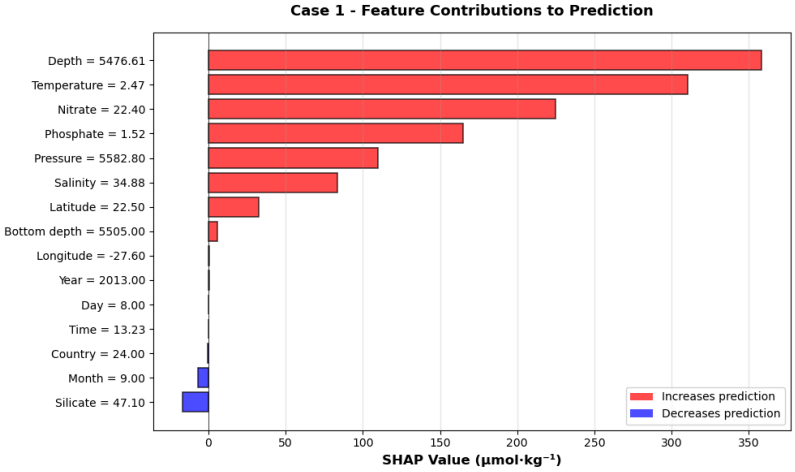
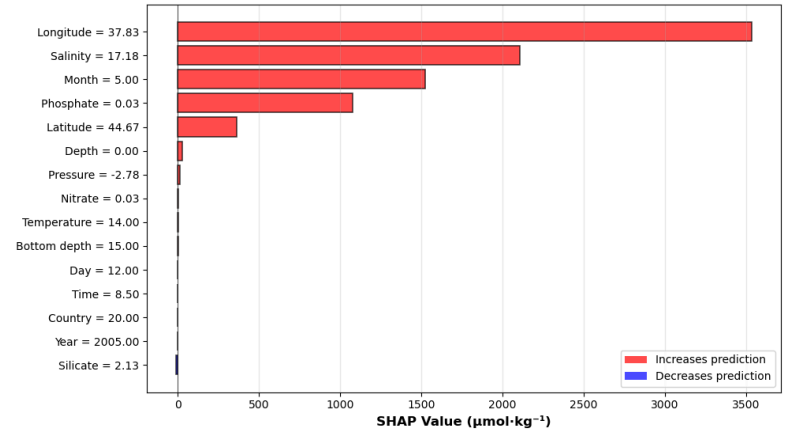


Figure C2: Residual Plots Showing Distribution of Predictions (0 line = perfect prediction, above = underprediction & below = overprediction) against Key Features with Three Cases of Prediction Failure Highlighted.



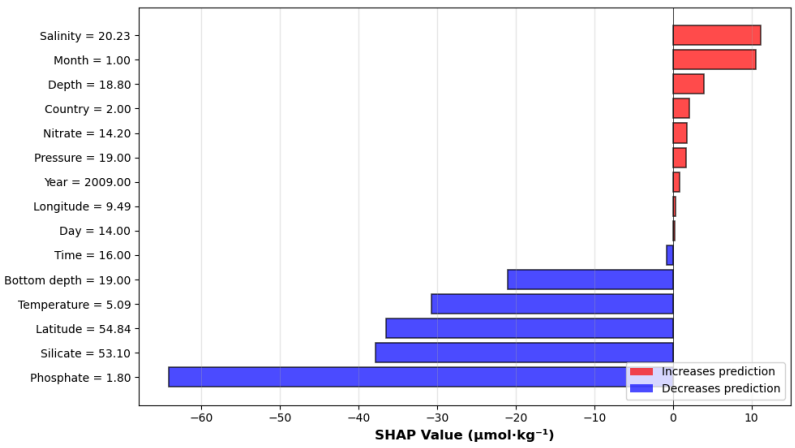
Feature Contributions:

	Feature	SHAP Value	Feature Value
8	Depth	358.286914	5476.6100
9	Temperature	310.240364	2.4722
13	Nitrate	224.699049	22.4000
11	Phosphate	164.967492	1.5200
14	Pressure	109.772657	5582.8000
10	Salinity	83.331698	34.8818
0	Latitude	32.473981	22.4997
12	Silicate	-16.917357	47.1000
3	Month	-6.591121	9.0000
7	Bottom depth	6.038776	5505.0000
1	Longitude	0.757851	-27.5992
2	Year	0.594644	2013.0000
6	Country	-0.357302	24.0000
4	Day	0.159371	8.0000
5	Time	0.062310	13.2300



Feature Contributions:

	Feature	SHAP Value	Feature Value
1	Longitude	3535.925107	37.833332
10	Salinity	2108.160657	17.180000
3	Month	1525.190971	5.000000
11	Phosphate	1078.962677	0.030820
0	Latitude	364.184203	44.666668
8	Depth	26.459232	0.000000
14	Pressure	13.100608	-2.775660
12	Silicate	-12.542230	2.129000
13	Nitrate	5.361961	0.031470
9	Temperature	5.009255	14.000000
7	Bottom depth	2.597280	15.000000
2	Year	-1.231385	2005.000000
6	Country	-0.471542	20.000000
5	Time	-0.345004	8.500000
4	Day	0.027036	12.000000



Feature Contributions:

	Feature	SHAP Value	Feature Value
11	Phosphate	-64.134606	1.800
12	Silicate	-37.785079	53.100
0	Latitude	-36.478227	54.840
9	Temperature	-30.732632	5.090
7	Bottom depth	-20.968476	19.000
10	Salinity	11.201767	20.230
3	Month	10.572756	1.000
8	Depth	3.972764	18.800
6	Country	2.081261	2.000
13	Nitrate	1.764781	14.200
14	Pressure	1.664677	19.000
2	Year	0.876192	2009.000
5	Time	-0.780190	16.000
1	Longitude	0.332951	9.495
4	Day	0.200886	14.000

Figure C3, C4 & C5: SHAP Plots showing Feature Contribution to Prediction of Three Cases of Prediction Failure

```
Distribution Check

Case 1 (Rank 1) (index: 468606)
--------------------------------------------------------------------------------

Feature Percentiles in Training Data:
  Depth       :  5476.61 ( 99.7%) <- extreme
  Salinity    :    34.88 ( 84.9%)
  Temperature :     2.47 ( 24.8%)
  Phosphate   :     1.52 ( 59.4%)

Case 2 (Rank 2) (index: 469811)
--------------------------------------------------------------------------------

Feature Percentiles in Training Data:
  Depth       :     2.77 (  8.3%)
  Salinity    :    32.74 ( 29.8%)
  Temperature :     4.25 ( 38.8%)
  Phosphate   :     0.42 ( 25.3%)

Case 3 (Rank 4) (index: 186984)
--------------------------------------------------------------------------------

Feature Percentiles in Training Data:
  Depth       :     0.00 (  0.0%) <- extreme
  Salinity    :    17.18 ( 20.3%)
  Temperature :    14.00 ( 81.8%)
  Phosphate   :     0.03 (  2.9%) <- extreme
```

```
Nearest Neighbor Analysis

Case 1 (Rank 1) (index: 468606)
--------------------------------------------------------------------------------
distance to nearest:  0.146
median to all train:  7.578

nearest 5 oxygen values: [241.8 242.7 242.1 243.8 242.6]
mean: 242.6, actual: 242.9

Case 2 (Rank 2) (index: 469811)
--------------------------------------------------------------------------------
distance to nearest:  0.365
median to all train:  4.565

nearest 5 oxygen values: [313.3 318.9 312.1 297.8 321.1]
mean: 312.6, actual: 314.8

Case 3 (Rank 4) (index: 186984)
--------------------------------------------------------------------------------
distance to nearest:  0.039
median to all train:  4.213

nearest 5 oxygen values: [9261.2 9349.1 9221.  9285.3  261. ]
mean: 7475.5, actual: 9573.5
```

Table C1 & C2: Feature Distribution Analysis for Extreme Value Identification and Nearest Neighbour Analysis for Training Data Similarity

```
Comparison to Similar Cases

Case 1 (Rank 1) (index: 468606)
--------------------------------------------------------------------------------
similar cases (depth 5000–5500m): 834
  this error:     1244.3
  similar mean:   3.1
  similar median: 0.4
  → 406.1x worse

Case 2 (Rank 2) (index: 469811)
--------------------------------------------------------------------------------
similar cases (high latitude (>60°N), shallow (<50m)): 4242
  this error:     1014.9
  similar mean:   8.7
  similar median: 4.1
  → 117.3x worse

Case 3 (Rank 4) (index: 186984)
--------------------------------------------------------------------------------
similar cases (salinity 15–20): 2323
  this error:     703.4
  similar mean:   14.5
  similar median: 7.5
  → 48.5x worse
```

```
Oxygen Violations at Shallow Depths

Shallow Records (<200m): 346,311
Percentage of Total Data: 57.4%
Shallow Oxygen > 450 (impossible): 3,889 (1.12%)
Shallow Oxygen > 400 (unlikely): 17,268 (4.99%)

Top 10 Worst Shallow Violations:
--------------------------------------------------------------------------------
        Oxygen  Depth  Salinity  Temperature Country  Year
186984  9573.5   0.00   17.1800      14.0000      RU  2005
186983  9349.1   0.00   17.2000      14.0000      RU  2005
186978  9285.3   0.00   16.4900      14.0000      RU  2005
186985  9261.2   0.00   17.2300      14.0000      RU  2005
186988  9221.0   0.00   17.2000      14.0000      RU  2005
388948  4222.0   2.97   31.5836       4.8293      US  2010
452136   743.3   1.00   11.7760       5.6980      DK  2013
452189   732.0   1.00    8.0700       3.6530      DK  2013
452188   719.8   3.00   13.8160       3.2590      DK  2013
450229   714.5   1.00   27.4200      -0.4900      DK  2013
```

Table C3 & C4: Comparison to Similar Feature Cases and Oxygen Violations Occurring at Shallow Depths

Appendix D

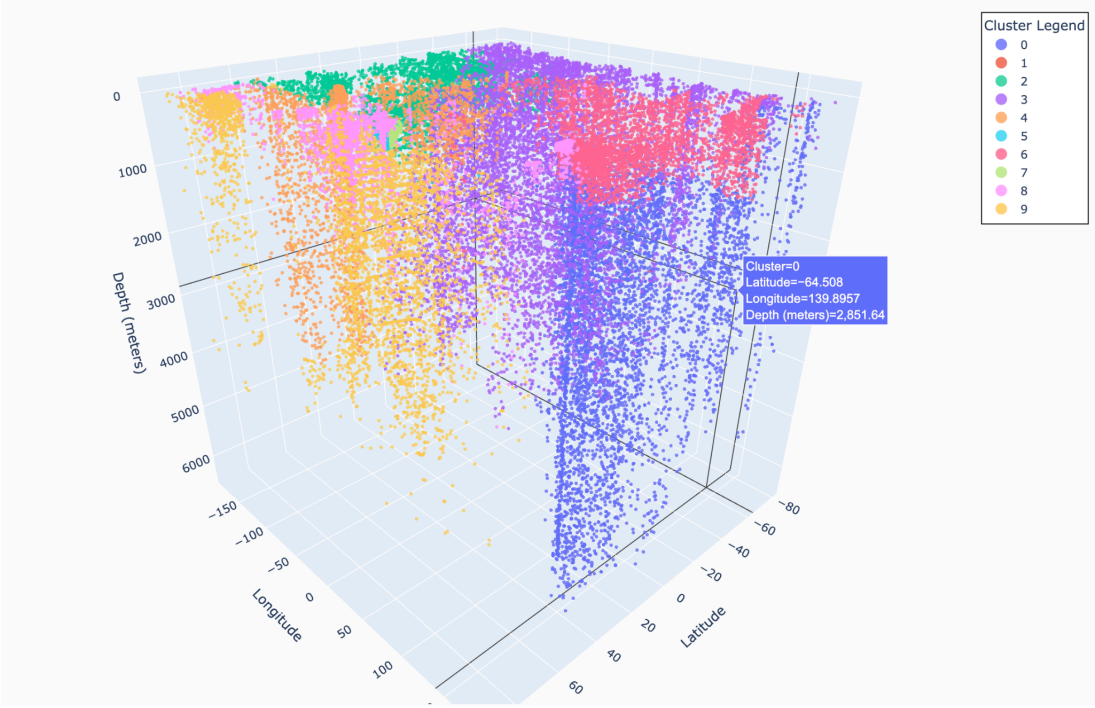


Figure D1: 3D Spatial Distribution of Clusters

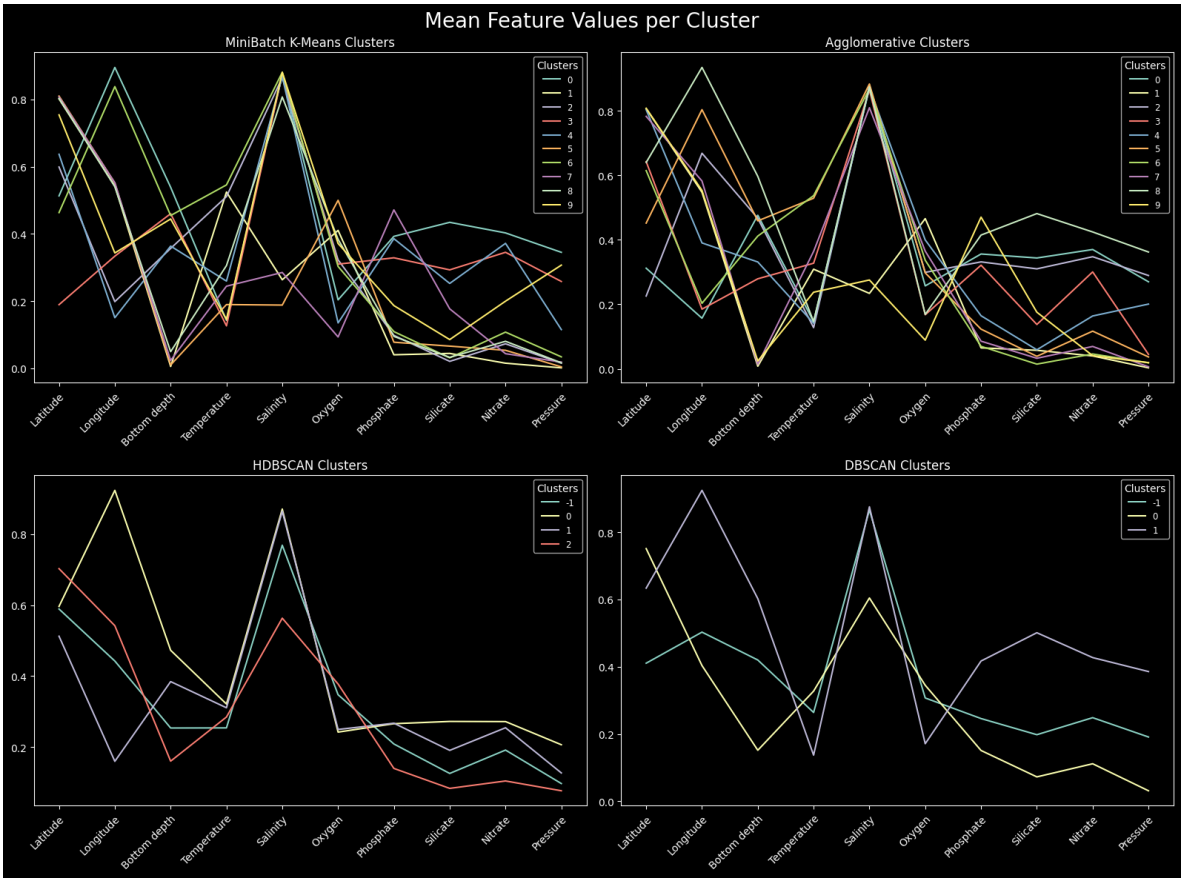


Figure D2: Line Plot of Mean Feature Values by Cluster