Remote Ocean Health Monitoring

SIADS 699 Capstone Project Report

Natasha Soldin (nsoldin), Ryan Mansfield (mansfire), Dharshana Somasunderam (sdharsha)

Introduction and Problem Statement

Ocean health is critical to global ecosystem functioning, climate regulation, and human livelihoods [1]. Dissolved oxygen (DO) is one of the most reliable indicators of ocean health, as it directly affects marine biodiversity, habitat suitability, and biogeochemical processes. Regions with depleted oxygen, or hypoxic zones, are uninhabitable for most marine life, leading to ecological and economic losses [2].

Monitoring DO concentrations is essential for tracking ocean health and detecting early signs of environmental degradation. Currently, DO monitoring relies on in-situ measurements collected by research vessels and autonomous instruments. While accurate, these methods are spatially and temporally sparse due to high operational costs, leaving the vast majority of the ocean unsampled and limiting our ability to detect emerging hypoxic events or assess global deoxygenation trends.

Satellite remote sensing offers a potential complement to sparse in-situ monitoring. Ocean colour satellites such as NASA's MODIS-Aqua [3,4] provide near-daily global coverage of surface ocean properties. While satellites cannot directly measure DO, they observe related conditions: chlorophyll-a concentration (indicating phytoplankton photosynthetic oxygen production) and sea surface temperature (controlling oxygen solubility and biological activity). These relationships suggest satellite data may be useful for predicting DO concentrations.

This project investigates whether satellite remote sensing data can predict surface ocean DO levels. In-situ oxygen measurements from the NOAA World Ocean Database [5,6] were integrated with coincident MODIS-Aqua satellite observations. Using supervised machine learning, predictive models were developed to assess the feasibility and accuracy of satellite-based oxygen estimation. The research addresses two main questions: (1) Can MODIS satellite data reliably predict DO concentrations in surface ocean waters? And (2) Which MODIS spectral bands and derived products are most predictive of oxygen levels?

If satellite-based oxygen prediction proves viable, it could provide continuous global monitoring to complement existing observing systems. This would improve the ability to track ocean deoxygenation, detect emerging hypoxic events, and support evidence-based policy decisions for marine conservation.

Literature Review

DO presents a fundamental challenge for satellite remote sensing: as a non-optically active variable, it does not directly influence the spectral parameters captured by satellite sensors [7]. Consequently, DO cannot be measured directly from remote sensing but must rather be estimated indirectly through correlations with optically active parameters such as sea surface temperature (SST) and chlorophyll-a concentration [8]. This limitation has driven the

development of statistical and machine learning approaches that leverage satellite-derived environmental variables to predict DO concentrations across diverse oceanographic climates.

Regional studies have demonstrated the viability of satellite-based DO estimates. Kim et al. applied multiple regression to MODIS and VIIRS data in Korean coastal waters, achieving 89.2% accuracy using SST and chlorophyll-a as predictors, and successfully detected summer deoxygenation trends in the Yellow Sea [7]. Song et al. employed Random Forest to map hypoxic zones in the northern Gulf of Mexico, identifying time lags of 0-19 days between surface processes and bottom water hypoxia [8]. Recent studies have employed more sophisticated methods: Li et al. achieved $R^2 = 0.958$ for three-dimensional DO reconstruction in the Mediterranean Sea using LightGBM combined with Biogeochemical-Argo float data [9], while Liu et al. demonstrated XGBoost's superior performance over Random Forest and Support Vector Regressor (SVR) for DO estimation in Lake Taihu using MODIS-derived temperature, chlorophyll-a, and Secchi depth (RMSE = 1.28 mg/L) [10].

Across these studies, temperature-related variables consistently emerged as the dominant predictors of DO variability, reflecting the well-established inverse relationship between temperature and oxygen solubility [11]. Performance metrics typically range from $R^2 = 0.6$-0.85 for regional models, with ensemble methods outperforming traditional regression approaches. However, existing research focuses predominantly on regional scales - coastal zones, semi-enclosed seas, or lakes - leaving a gap in global-scale satellite-based DO prediction. This project addresses that gap by integrating NOAA World Ocean Database measurements with MODIS-Aqua observation across diverse ocean regions spanning 2002-2023.

Methodology

Data Sources

This project integrated two primary data sources: in-situ dissolved oxygen (DO) measurements and remote sensing satellite observations.

The in-situ DO measurements were obtained from the NOAA World Ocean Database (WOD), specifically the Ocean Station Data (OSD) subset. The OSD contains bottle samples and low-resolution Conductivity-Temperature-Depth (CTD) measurements collected from oceanographic research vessels globally. The WOD structures data as oceanographic casts, which are vertical profiles of measurements taken at discrete depths as instruments descend through a water column. Data were accessed via the WOD Select Interface (https://www.ncei.noaa.gov/access/world-ocean-database-select/dbsearch.html) and downloaded in CSV format. The data were temporally filtered between 1990-2023 to provide adequate sample size and balance historical coverage with data quality improvements in modern oceanographic sampling. The key variables that were extracted are in Table 1 below and the complete OSD data schema is detailed in Appendix A.

Variable Name	Description	Unit
Cast ID	Unique Cast Identifier	-
Latitude, Longitude	Geographic co-ordinates	°
Date	Observation Date	Year-Month-Day
Oxygen	Dissolved Oxygen Concentration	µmol/kg
Depth	Depth of Captured Measurement	m

Table 1: Extracted NOAA WOD Dataset

Satellite remote sensing data were obtained from NASA's MODIS-Aqua Level-3 Standard Mapped Image (L3SMI) product via Google Earth Engine (GEE). MODIS-Aqua provides global ocean colour measurements at 4km spatial resolution and has a temporal resolution of 1-2 days, providing near-daily coverage. This dataset was accessed using GEE's API and corresponding date-coordinate pairs from the NOAA WOD dataset. The key variables that were extracted are in Table 2 below and the complete MODIS data schema is detailed in Appendix B.

Variable Name	Description	Unit
Latitude, Longitude	Geographic co-ordinates	°
Date	Observation Date	Year-Month-Day
chlor_a	Chlorophyll-a concentration	mg/m^3
nflh	Normalized fluorescence line height	mW cm^-2 µm^-1 sr^-1
poc	Particulate organic carbon	mg/m^3
Rrs (10 bands)	Remote sensing reflectance at bands: 412, 443, 469, 488, 531, 547, 555, 645, 667, 678 nm	sr^-1
sst	Sea surface temperature	°C

Table 2: Extracted MODIS-Aqua Dataset

Data Processing

Figure 1 below details the complete data processing workflow, showing the transformation from raw datasets to a combined and machine learning ready dataset.

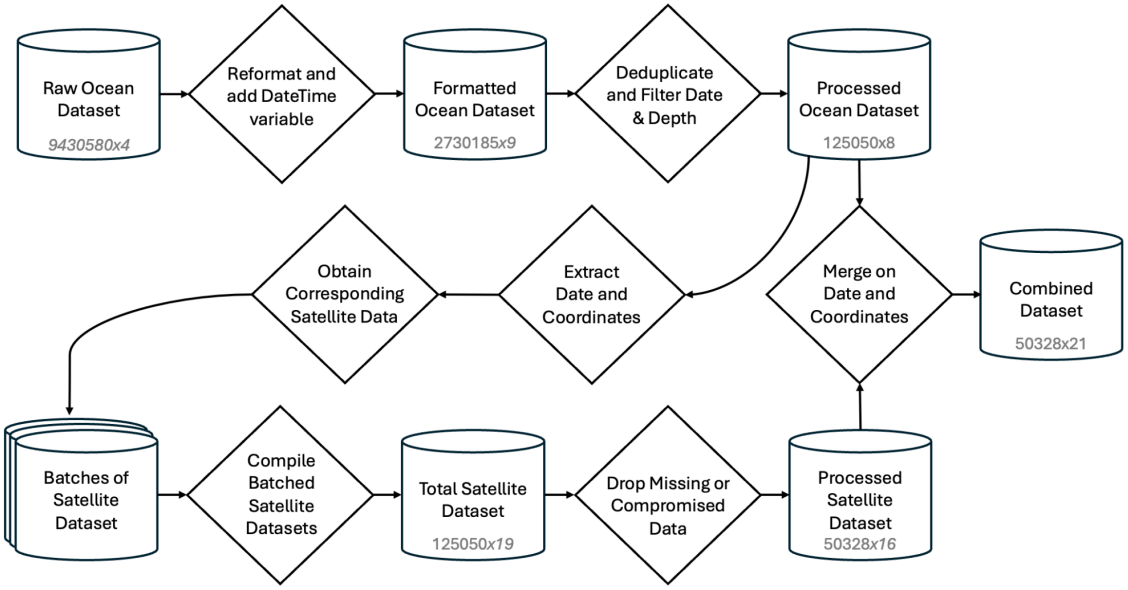


Figure 1: Unified Modeling Language (UML) Data Flow Diagram of the data processing pipeline showing integration of in-situ ocean and satellite datasets. Dataset dimensions (rows x columns) are shown at each stage.

The raw ocean dataset contained over 2.7 million oxygen measurements spanning multiple decades. To align with MODIS-Aqua's operational period (July 2002 onwards) and ensure compatibility with satellite observations, three key filtering steps were applied: temporal filtering, duplicate removal (likely present due to multiple vessels sampling the same location or data being submitted multiple times), and retention of only surface measurements. Satellites can only observe the ocean's uppermost layer, so deeper measurements were excluded to align with this penetration depth. The depth variable was then dropped to ensure models rely solely on satellite-observable features - a requirement for operational applications where ship-based depth measurements are unavailable. This produced 125,050 surface oxygen observations.

For each ocean measurement, corresponding satellite data were extracted using a ±3 day temporal window [12] - a standard approach in ocean color validation that balances temporal precision with cloud-free data availability. However, satellite coverage is inherently incomplete due to cloud interference and sensor limitations. Of the 125,050 matched observations, only 50,328 (40.2%) had complete satellite data across all required variables. Missing values in sea surface temperature (1.1%) and particulate organic carbon (1.3%) were imputed using spatial averaging and regression techniques respectively.

The final integrated dataset contained 50,328 paired ocean-satellite observations spanning 2002–2023. The dataset was split into training (70%, n=35,230), validation (15%, n=7,549), and testing (15%, n=7,549) subsets for model development and evaluation.

Feature Engineering

Following data processing and integration, the final dataset underwent feature engineering to transform raw satellite measurements into features optimized for predictive modeling, with decisions guided by exploratory data analysis detailed in Appendix C. Candidate features were generated across seven categories: temperature transformations, log transformations of skewed variables, biomass ratios, spectral band ratios, normalized indices, interaction terms, and spatial-temporal features.

Feature selection was performed using Pearson correlation analysis between candidate features and dissolved oxygen in the training dataset. A correlation threshold of $|r| > 0.4$ was applied, retaining nine numerical features. To address multicollinearity, features exhibiting near-perfect correlation ($|r| > 0.95$) with other selected features were removed: sst_cubed ($r = 0.982$ with sst_squared), ratio_443_555 ($r = 0.998$ with ratio_443_547), and log_poc ($r = 0.950$ with log_chlor_a). This yielded a final set of six numerical features (sst_squared, sst_anomaly, abs_latitude, log_chlor_a, ratio_443_547, sst_chlor_interaction) and the categorical Season variable. Complete candidate features and their correlations are provided in Appendix D. The final modeling dataset comprised 13 original satellite variables, 7 engineered features, 5 metadata fields, and the target variable oxygen.

Modelling Approach

Five models representing distinct machine learning families were selected as baseline models: Linear Regression, Random Forest, XGBoost, LightGBM, and a Multilayer Perceptron Neural Network. This selection spans linear, tree-based ensemble, and neural network approaches to assess which architecture best captures oceanographic patterns. For advanced models, Linear Regression was replaced with Elastic Net, which combines L1 and L2 regularization for improved handling of correlated features.

Hyperparameter tuning using RandomizedSearchCV was implemented on each advanced model. Tuned parameters included regularization strength, tree depth, learning rate, and architecture-specific settings. Detailed search spaces and optimal configurations are presented in Appendix E. Cross-validation employed a 5-fold TimeSeriesSplit method, which preserves temporal ordering by training each fold exclusively on past observations before validating on subsequent ones.

Data Leakage Prevention

Several measures were implemented throughout the analysis pipeline to prevent data leakage. The dataset was split temporally into training (70%, 2002-2012), validation (15%, 2012-2015), and testing (15%, 2015-2023) sets. Temporal splitting avoids leakage from autocorrelated ocean measurements where adjacent observations share similar conditions. All preprocessing operations - including exploratory data analysis, feature engineering and feature scaling - were performed exclusively on training data and applied to validation and test sets. For any cross-validation procedures, TimeSeriesSplit was employed instead of K-Fold to respect temporal ordering. The validation set was used for model development and hyperparameter tuning, while the test set was reserved for final evaluation only.

Results

Baseline Models

Models were evaluated on Coefficient of Determination (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). R^2 measures the proportion of oxygen variance explained by predictor variables. MAE and RMSE both quantify prediction error but differ in sensitivity: MAE provides robust assessment across the dataset, while RMSE emphasizes larger errors and outlier sensitivity.

Model	R^2	MAE	RMSE
Linear Regression	0.588	23.333	39.293
Random Forest	0.965	5.601	11.505
LightGBM	0.793	15.842	27.843
XGBoost	0.874	12.640	27.722
Neural Network	0.807	15.715	26.908

Table 3: Performance Metrics of Baseline Models applied to Training Data

Model	R^2	MAE	RMSE
Linear Regression	0.537	26.717	43.611
Random Forest	0.5872	23.7115	40.248
LightGBM	0.596	23.485	39.824
XGBoost	0.595	24.017	39.889
Neural Network	0.158	39.979	57.472

Table 4: Performance Metrics of Baseline Models applied to Validation Data

Baseline models show moderate validation performance, with the highest R^2 of 0.596 achieved by the LightGBM model, slightly outperforming the XGBoost model (R^2 = 0.595). All models exhibited high prediction errors (MAE = 23.48 - 39.98 µmol/kg, RMSE = 39.82 - 57.47 µmol/kg) and experienced performance gaps between training and validation sets, indicating overfitting. This was most pronounced in the Neural Network, which dropped from R^2 = 0.807 (training) to R^2 = 0.158 (validation).

Advanced Models

Model	R^2	MAE	RMSE
Elastic Net	0.586	23.568	39.381

Random Forest	0.898	8.856	19.525
LightGBM	0.710	18.160	32.948
XGBoost	0.801	15.486	27.312
Neural Network	0.666	19.583	35.388

Table 5: Performance Metrics of Advanced Models applied to Training Data

Model	R^2	MAE	RMSE
Elastic Net	0.535	27.007	42.718
Random Forest	0.593	23.176	39.969
LightGBM	0.576	24.112	40.787
XGBoost	0.592	23.639	39.993
Neural Network	0.567	25.215	41.243

Table 6: Performance Metrics of Advanced Models applied to Validation Data

Hyperparameter tuning yielded minimal performance changes, with most models showing slight decreases in validation R^2. Random Forest achieved the highest validation R^2 (0.593), closely followed by XGBoost (0.592). However, Random Forest exhibited substantial overfitting (training R^2 = 0.898, validation R^2 = 0.593) and had a longer training time to achieve similar results.

Best Model Selection and Final Evaluation

XGBoost was selected as the final model based on its optimal balance between validation performance (R^2 = 0.592), relatively low prediction errors (MAE = 23.639 µmol/kg, RMSE = 39.993 µmol/kg), and lower computational requirements and superior generalization compared to Random Forest (training-validation gap: 0.209 vs. 0.305). Comprehensive diagnostics were then performed on the selected XGBoost model, including feature importance analysis, ablation study, and trend analysis (Appendix F). Feature importance analysis further validated XGBoost's selection: it utilized balanced feature contributions (Figure 3), while Random Forest relied heavily on geographic coordinates (57% spatial), indicating potential overfitting to location.

The final tuned XGBoost configuration (Appendix E) was evaluated on the held-out test set (2015-2023) to assess generalization to unseen data (Table 7; Figure 2).

Model	R^2	MAE	RMSE
XGBoost	0.728	18.908	29.863

Table 7: Performance Metrics of Best Performing Model applied to Testing Data

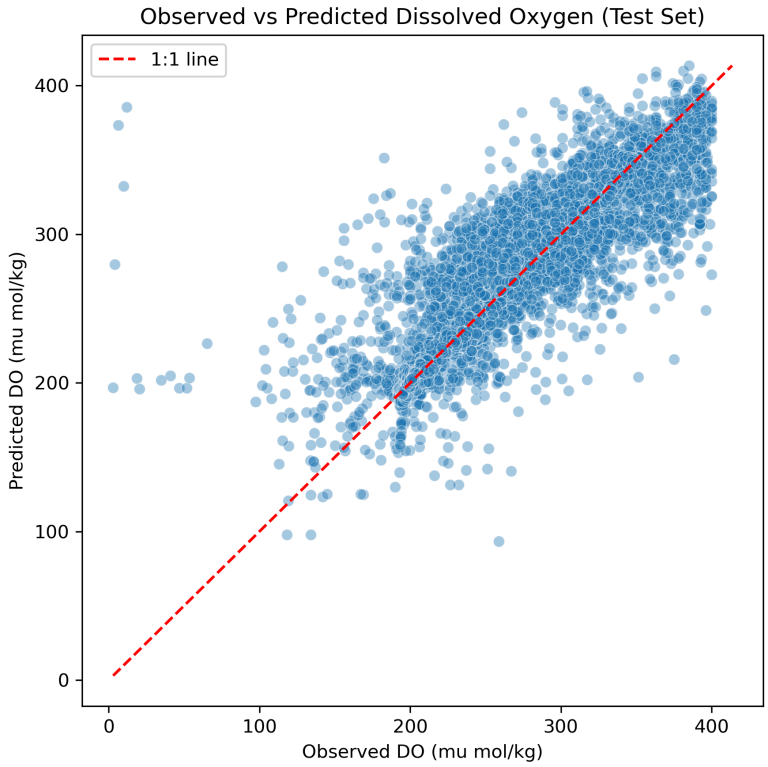


Figure 2: Observed vs Predicted Dissolved Oxygen Concentrations (µmol/kg) for the Final Tuned XGBoost model on the test set. Dashed red line indicates perfect prediction (1:1)

Discussion

Feature Importance

Feature importance analysis (Figure 3) revealed that spatial variables (34%), temperature-related features (20%), and seasonal indicators (16%) were the primary predictors of dissolved oxygen. These features are fundamentally interconnected through temperature: latitude captures geographic temperature gradients (warm equator, cold poles), SST measures it directly, and seasonal patterns capture temporal variation. This suggests that the model primarily learned the temperature-oxygen solubility relationship - warmer water holds less dissolved oxygen due to reduced gas solubility [1] - rather than complex oceanographic processes. Notably, biological indicators such as chlorophyll-a contributed minimally (<5%), further emphasising the model's reliance on temperature proxies rather than capturing genuine oxygen production and consumption dynamics.

Model Performance and Generalization

The final XGBoost model achieved a test R^2 of 0.728, exceeding validation performance (R^2 = 0.592). This unexpected improvement may reflect differences in oceanographic conditions between time periods - the test period (2015-2023) potentially exhibiting more consistent patterns than the validation period (2012-2015). Minimal geographic overlap between training and test sets (8.6%) suggests the model learned generalizable spatiotemporal patterns rather

than memorizing specific locations. The strong test performance validated XGBoost's selection over Random Forest, demonstrating that lower overfitting led to effective generalization.

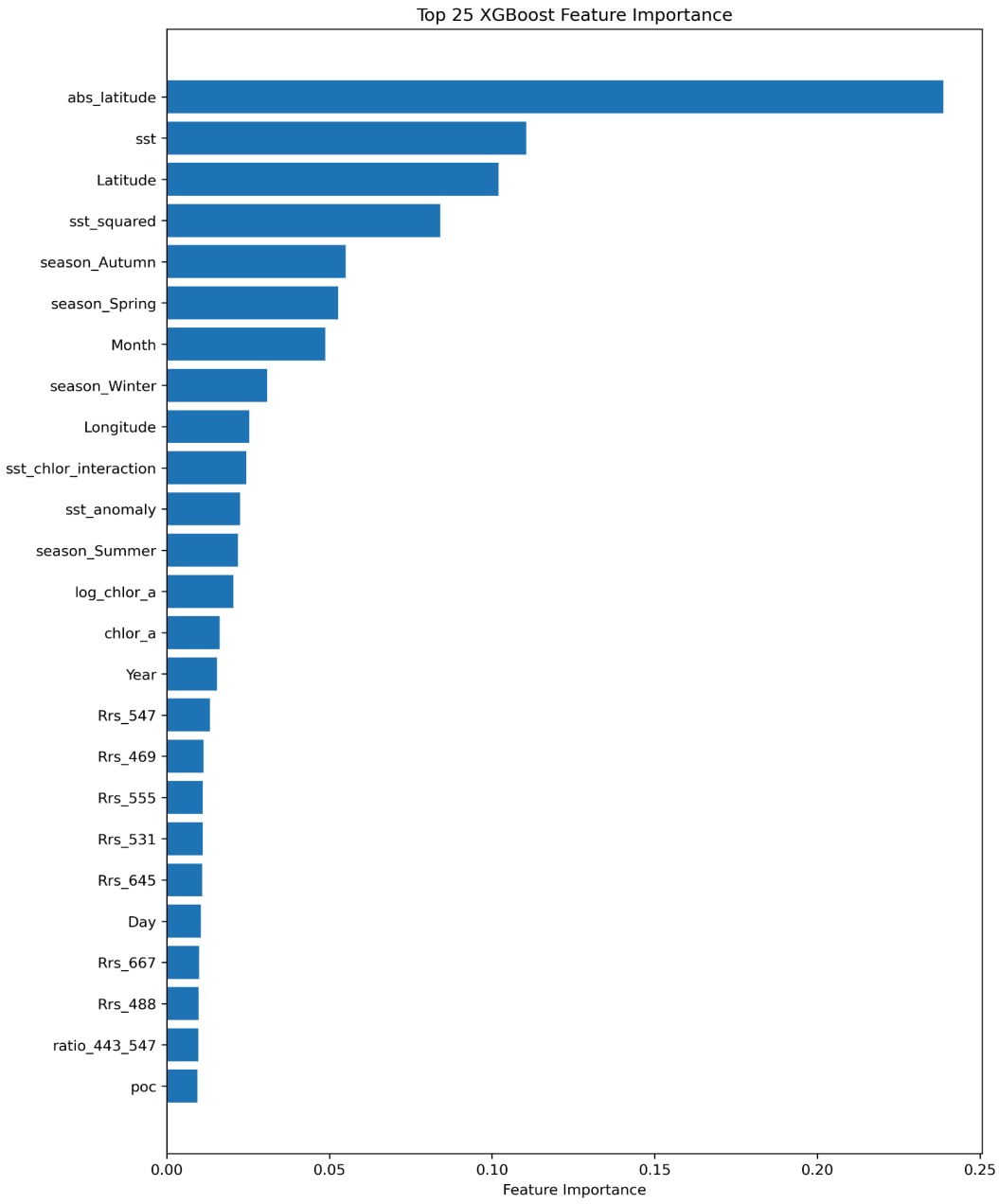


Figure 3: Feature Importance Rankings for the Final Tuned XGBoost Model

Model Limitations

Despite satisfactory test performance (R^2 = 0.728, RMSE = 29.86 µmol/kg, MAE = 18.91 µmol/kg), the model exhibited several important limitations. A consistent validation R^2 ceiling near 0.59 emerged across all model architectures, with hyperparameter tuning yielding negligible improvements ($\Delta R^2 < 0.03$). This performance plateau indicates a fundamental predictability limit with the current feature set rather than suboptimal model selection,

suggesting that architectural refinements alone have limited impact. This ceiling likely reflects several constraints: (1) fundamental limitations of inferring subsurface oxygen dynamics from surface observations, as satellites cannot directly measure critical processes such as vertical mixing, nutrient dynamics, and biological oxygen consumption rates; and (2) substantial natural variability across diverse ocean regions and climatic conditions spanning 20 years. As identified in the feature importance analysis, the model's reliance on temperature-related variables suggests it captured thermal solubility patterns but not the biological and physical processes that drive oxygen variability independent of temperature.

Additional limitations include sensitivity to outliers, evidenced by elevated RMSE relative to MAE, and failure in hypoxia detection (DO < 60 μmol/kg: precision, recall, F1 = 0.000 - Figure G1, Appendix G). The severe class imbalance - only 12 hypoxic samples among 7,103 test observations - prevented the model from learning patterns associated with rare hypoxic events.

Future Work

The model's consistent validation performance ceiling and reliance on temperature-related dynamics for oxygen prediction suggests that satellite surface observations alone cannot fully capture dissolved oxygen variability. Improving predictions will likely require data on subsurface processes - vertical mixing, nutrient concentrations and biological activity - currently invisible to satellites. Hybrid models combining satellite data with in-situ subsurface measurements offer a potential path forward. Region-specific models for distinct ocean zones and targeted approaches for hypoxia detection using class-balancing techniques may also improve performance.

Ethical Considerations and Broader Impact

If developed further, this model could support early detection of declining ocean health by identifying regions at risk of oxygen depletion, providing data-driven evidence for policy recommendations. The approach could also inspire similar satellite-based monitoring systems for other environmental indicators such as rainforest health, pollution tracking or ice cap loss.

However, several ethical concerns warrant attention. The model's failure in hypoxia detection limits its ability to warn of the low-oxygen events most threatening to marine ecosystems and fishing-dependent communities. Additionally, sampling bias toward the Northern Hemisphere (Figure C3, Appendix C) may reduce accuracy in the global south - regions often most dependent on marine resources - potentially exacerbating environmental inequalities. Finally, model outputs could be misused: fishing companies might exploit predictions to target oxygen-rich areas, or stakeholders could selectively present findings to justify harmful developments. Future research should prioritize data collection in underrepresented regions, and transparent governance with equitable access will be essential to prevent misuse.

Conclusion

This study demonstrated the viability of predicting surface ocean dissolved oxygen from satellite data using machine learning, with XGBoost achieving test R^2 of 0.728 on geographically distinct locations. The model successfully learned generalizable oceanographic patterns, particularly

latitudinal oxygen gradients driven by temperature. However, fundamental limitations emerged. All models reached a validation R^2 ceiling near 0.59 regardless of architecture or tuning, indicating constraints imposed by satellite-derived features rather than model design. The reliance on spatial variables and complete hypoxia detection failure highlight that predicting subsurface dynamics from surface observations alone faces inherent physical constraints. Nevertheless, the demonstrated predictive skill for surface oxygen patterns suggests satellite-based approaches can complement - though at this stage not replace - traditional in-situ monitoring networks.

Statement of Work

- Natasha Soldin: Data acquisition, processing, missing value handling and merging, exploratory data analysis, feature engineering, code review and report writing: introduction and problem statement, methodology, conclusion, appendices A-D
- Ryan Mansfield: Model development and optimisation (baseline and advanced models), advanced model evaluation (feature importance and ablation analyses), feature engineering and report writing: methodology, results, discussion, appendix E-F
- Dharshana Somasunderam: Final model evaluation & appendix G

References

[1] United Nations (2021). The Second World Ocean Assessment: World Ocean Assessment II. Cambridge University Press. https://doi.org/10.18356/9789216040062

[2] Breitburg, D., Levin, L. A., Oschlies, A., Grégoire, M., Chavez, F. P., Conley, D. J., Garçon, V., Gilbert, D., Gutiérrez, D., Isensee, K., Jacinto, G. S., Limburg, K. E., Montes, I., Naqvi, S. W. A., Pitcher, G. C., Rabalais, N. N., Roman, M. R., Rose, K. A., Seibel, B. A., Telszewski, M., Yasuhara, M., & Zhang, J. (2018). Declining oxygen in the global ocean and coastal waters. *Science*, *359*(6371), eaam7240. https://doi.org/10.1126/science.aam7240

[3] NASA Goddard Space Flight Center, Ocean Ecology Laboratory, Ocean Biology Processing Group. (2018). *Moderate-resolution Imaging Spectroradiometer (MODIS) Aqua Ocean Color Data*. NASA OB.DAAC. Greenbelt, MD, USA. Accessed via Google Earth Engine. https://developers.google.com/earth-engine/datasets/catalog/NASA_OCEANDATA_MODIS-Aqua_L3SMI

[4] NASA Earthdata. (2021). Data Processing Levels. [online] Available at: https://www.earthdata.nasa.gov/learn/earth-observation-data-basics/data-processing-levels

[5] Mishonov A.V., T. P. Boyer, O. K. Baranova, C. N. Bouchard, S. Cross, H. E. Garcia, R. A. Locarnini, C. R. Paver, J. R. Reagan, Z. Wang, D. Seidov, A. I. Grodsky, J. G. Beauchamp, (2024): World Ocean Database 2023. C. Bouchard, Technical Ed., NOAA Atlas NESDIS 97, 206 pp., https://doi.org/10.25923/z885-h264

[6] Garcia, H. E., Boyer, T. P., Locarnini, R. A., Reagan, J. R., Mishonov, A. V., Baranova, O. K., Paver, C. R., Wang, Z., Bouchard, C., Cross, S., Seidov, D., & Dukhovskoy, D. (2024). *World Ocean Database 2023 User's Manual*. NOAA Atlas NESDIS 98. National Centers for Environmental Information. https://doi.org/10.25923/j8gq-ee82

[7] Kim, Y.H., Son, S., Kim, H.C., Kim, B., Park, Y.G., Nam, J., & Ryu, J. (2020). Application of satellite remote sensing in monitoring dissolved oxygen variabilities: A case study for coastal waters in Korea. *Environment International*, 134, 105301. https://doi.org/10.1016/j.envint.2019.105301

[8] Li, Y., Robinson, S.V.J., Nguyen, L.H., & Liu, J. (2023). Satellite prediction of coastal hypoxia in the northern Gulf of Mexico. *Remote Sensing of Environment*, 284, 113346. https://doi.org/10.1016/j.rse.2022.113346

[9] Liu, G., Yu, X., Zhang, J., & Wang, X. (2025). Reconstruction of the three-dimensional dissolved oxygen and its spatio-temporal variations in the Mediterranean Sea using machine learning. *Journal of Environmental Sciences*, 157, 710–728. https://doi.org/10.1016/j.jes.2025.01.010

[10] Liu, M., Wang, L., & Qiu, F. (2022). Using MODIS data to track the long-term variations of dissolved oxygen in Lake Taihu. *Frontiers in Environmental Science*, 10, 1096843. https://doi.org/10.3389/fenvs.2022.1096843

[11] Mahaffey, C., Palmer, M., Greenwood, N., & Sharples, J. (2020). Impacts of climate change on dissolved oxygen concentration relevant to the coastal and marine environment around the UK. MCCIP Science Review 2020, 31-53. https://doi.org/10.14465/2020.arc02.oxy

[12] Lai, L., Hou, X., Han, W., & Duan, H. (2025). Concerns about the temporal matching windows in satellite-ground synchronization for lacustrine environment mapping. *Water Research*, 286, 124208. https://doi.org/10.1016/j.watres.2025.124208

Appendix A

The following appendix provides an overview of the NOAA World Ocean Database (WOD) structure. The WOD contains 11 different datasets organized by instrument type (OSD, CTD, XBT, MBT, SUR, APB, MRB, PFL, DRB, UOR, GLD), all following the same seven-component data structure described below. This project specifically uses the Ocean Station Data (OSD) subset, which contains bottle samples and low-resolution CTD measurements.

1. Primary Header: this header is present for every cast and provides fundamental metadata necessary to locate and identify a given observation in space and time. Core identifiers are listed in Table A1 below.

Variable Name	Description	Format/ Unit
WOD Version Identifier	WOD release version	Character
WOD Unique Cast Number	Unique identifier for each cast	Integer
Country Code	ISO country code	2-Character
Cruise Number	Cruise Identification Number	Integer
Year	Year of observation	Integer (YYYY)
Month	Month of observation	Integer (MM)
Day	Day of observation	Integer (DD)
Time	Time of observation	Decimal (hrs)
Latitude	Latitude of observation	Decimal (°)
Longitude	Longitude of observation	Decimal (°)
Number of Depth Levels	Count of depth measurements on cast	Integer
Station Type	Observed (0) or Standard (1) level data	Integer
Number of Variables	Count of measured Variables	Integer
Variable Codes	List of variable codes present	Integer array

Table A1: Primary Header Variables

2. Secondary Header: this header includes 70+ optional metadata fields, with not every field populated for every cast (this is dependent on data collection platform, program and historical period). This includes meteorological observations, instrument information, data provenance, sampling details and quality indicators. *For a complete listing, see Tables 4a and 4b of the WOD23 User's Manual* [6].

3. Variable Specific Secondary Header: this includes metadata explicitly tied to each measured variable in the cast, including accession numbers, project associations, measurement scales, protocols and units. *For a complete listing, see Tables 5a and 5b of the WOD23 User's Manual* [6].
4. Character Data: text-based identifiers that supplement numeric codes including originator's cruise identifier, station code and principal investigator code.
5. Biological Header: sampling metadata for plankton observations including net type, mesh size, volume samples and collection methods. *For a complete listing, see Tables 6a and 6b of the WOD23 User's Manual* [6].
6. Taxa-Specific and Biomass Data: this includes individual taxonomic group observations like WOD tax codes, concentration and total biomass measurements. *For a complete listing, see Tables 7a, 7b, 8 and 9 of the WOD23 User's Manual* [6].
7. Measured Variables: this includes all oceanographic measured variables taken at discrete depths. Table A2 below lists the variables available in the OSD dataset. *For a complete listing, see Tables 3a and 3b of the WOD23 User's Manual* [6].

Code	Variable Name	Unit
1	Temperature	Degrees Celsius (°C)
2	Salinity	Dimensionless
3	Oxygen	µmol/kg
4	Phosphate	µmol/kg
6	Silicate	µmol/kg
8	Nitrate	µmol/kg
9	pH	Dimensionless
11	Chlorophyll	µg/l
17	Alkalinity	mmol/l
20	Partial Pressure of Carbon Dioxide	µatm
21	Dissolved Inorganic Carbon	mmol/l
24	Transmissivity	1/m
25	Pressure	Decibar
33	Tritium	Tritium Unit (TU)
34	Helium	nmol/kg

35	Helium-3	%
36	Carbon-14	Per mille; parts per thousand
37	Carbon-13	Per mille; parts per thousand
38	Argon	nmol/kg
39	Neon	nmol/kg
40	Chlorofluorocarbon-11	pmol/kg
41	Chlorofluorocarbon-12	pmol/kg
42	Chlorofluorocarbon-113	pmol/kg
43	Oxygen-18	Per mille; parts per thousand

Table A2: Depth-Dependent Measured Variables from the OSD Dataset

Appendix B

The following appendix provides an overview of NASA's MODIS database structure. The Moderate Resolution Imaging Spectroradiometer (MODIS) operates aboard two NASA satellites: Terra (launched 1999) and Aqua (launched 2002). Terra provides morning observations while Aqua provides afternoon observations, with the latter offering improved conditions for ocean color retrieval due to reduced sun glint and more stable atmospheric conditions. This project uses MODIS-Aqua data exclusively, accessed via Google Earth Engine [3], to leverage these advantages for ocean monitoring.

Specifically, the MODIS-Aqua Level-3 Standard Mapped Image (L3SMI) products, which provides near-daily global ocean color and thermal data at 4km spatial resolution from July 2002 onwards, were used for this project. MODIS data processing follows a hierarchical structure (Table B1) [4]. L3SMI products are gridded image representations using an Equidistant Cylindrical projection, stored in NetCDF4 format. These products provide numerical geophysical parameters derived from spectral measurements.

Level	Description
L0	Reconstructed, unprocessed instrument and payload data at full resolution with any form of communication artifacts removed
L1A	Reconstructed, unprocessed instrument data at full resolution, time-referenced and annotated
L1B	L1A data processed to instrument units
L1C	L1B data processed with additional variables to describe spectra
L2	Derived geophysical variables at the same resolution and locations as L1 source data
L2A	Information derived from geolocated instrument data (e.g. ground or surface return elevation)
L2B	L2A data processed to instrument units
L3	Variables mapped to uniform space-time grids with completeness and consistency
L3A	Periodic summaries of L2 products (e.g. weekly, 10-day, monthly)
L4	Model output or results from analyses of lower-level data

Table B1: NASA Data Processing Levels [4]

MODIS-Aqua L3SMI variables used in this project are described in Table 2 and represent the variables available via Google Earth Engine. The broader MODIS-Aqua product suite includes additional L3 ocean color datasets such as particulate inorganic carbon (PIC), diffuse attenuation coefficient (Kd_490), photosynthetically available radiation (PAR), and aerosol properties, distributed as separate files by NASA OB.DAAC but not included in the Google Earth Engine collection - a trade-off for the ease of access provided by the platform.

Appendix C

The following appendix details the Exploratory Data Analysis (EDA) performed on the training dataset prior to, and for the purpose of informing, the feature engineering section.

Distribution Analysis

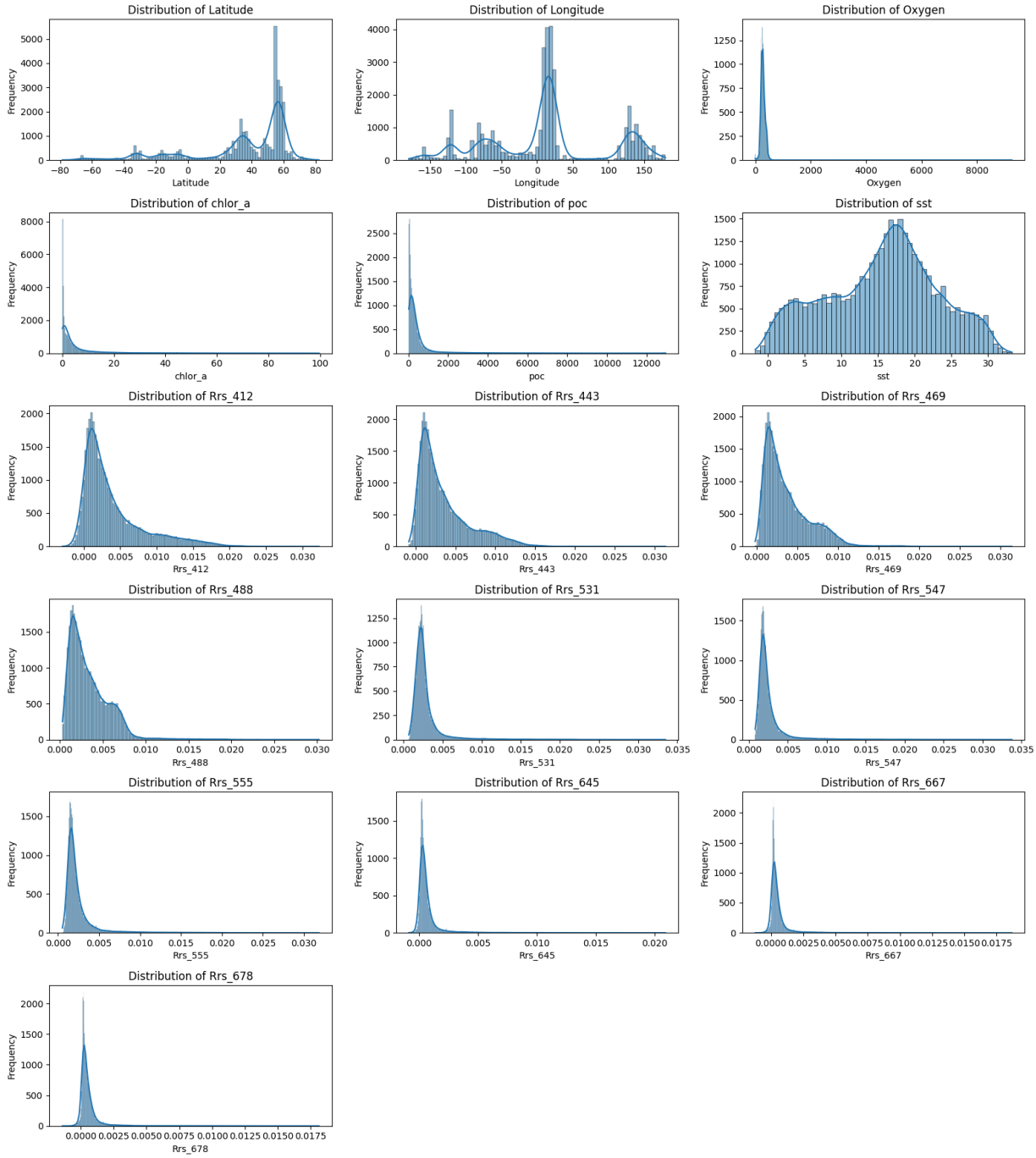


Figure C1: Histogram Distribution Plots of the Training Dataset's Numeric Variables

Most variables exhibit severe right-skewness; only SST shows normal distribution. Latitude and longitude display multimodal patterns reflecting discrete oceanographic sampling campaigns. These findings justify subsequent skewness analysis and geographic coverage assessment.

Skewness Analysis

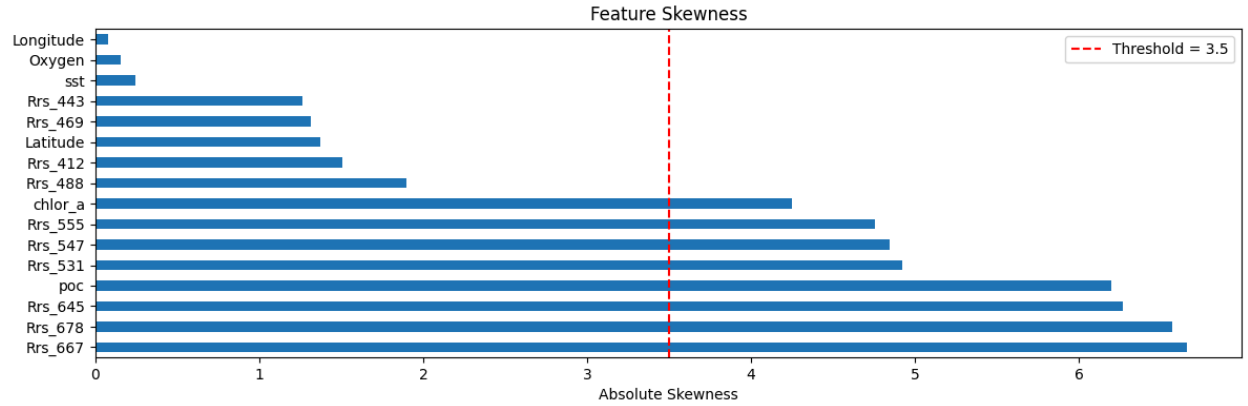


Figure C2: Feature distribution skewness analysis. Red line indicates |skewness| threshold.

Spectral reflectance bands exhibit the most severe skewness (Rrs_667: 6.8, Rrs_678: 6.7, Rrs_645: 6.3), followed by biological variables (POC: 6.4, chlor_a: 4.1). Features exceeding the skewness threshold of 3.5 were subjected to log transformations during feature engineering to evaluate whether transformed versions improved predictive performance.

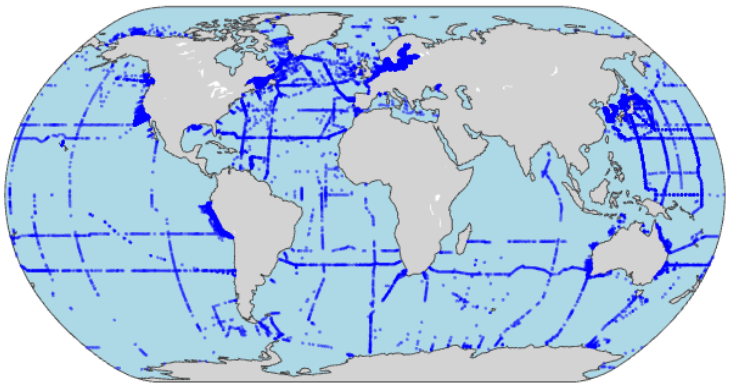
Oxygen Quality Control

Outlier analysis revealed biologically implausible oxygen values likely resulting from sensor calibration errors, measurement artifacts, or data entry mistakes. A total of 2,806 outliers (7.97%) exceeding 400 µmol/kg or below 0 µmol/kg [11] were removed from all splits, reducing the dataset to 32,423 training, 6,922 validation, and 7,103 testing samples. Future work should investigate the sources of these anomalies to improve data quality protocols. Following outlier removal, oxygen concentrations were categorized into biogeochemical zones (Table C1).

Category	Range (µmol/kg)	Count	%
Severe Hypoxia	< 22	88	0.3
Hypoxic	22 - 60	72	0.2
Low Oxygen	60 - 192	1,444	4.5
Normal	192 - 320	24,123	74.4
High Oxygen	> 320	6,696	20.7

Table C1: Distribution of Oxygen Measurements by Biogeochemical Category [11]

Geographic Coverage



The training dataset is fairly well distributed geographically but it is more concentrated in the Northern Hemisphere. Tropical and Southern Hemisphere regions show sparse coverage, limiting model generalizability to these areas. This geographical bias should be noted as a limitation.

Figure C3: Geographic Distribution of Training Measurements

Correlation Analysis

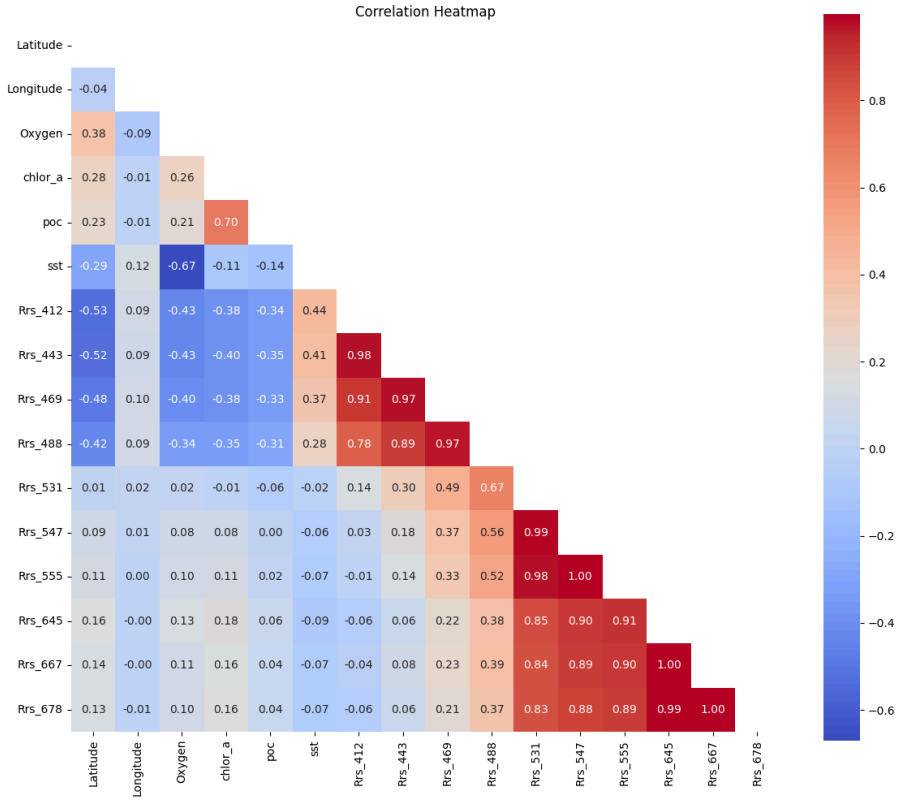


Figure C4: Pearson correlation matrix of numeric variables in the training dataset

SST shows the strongest correlation with oxygen (r = -0.67), followed by blue spectral reflectance bands (Rrs_412: r = -0.43, Rrs_443: r = -0.43, Rrs_469: r = -0.40). Latitude displays moderate positive correlation (r = 0.38), while biological variables show weaker relationships (chlor_a: r = 0.26, POC: r = 0.21).

Appendix D

The following Appendix details all candidate engineered features generated during the feature engineering process and their Pearson correlation coefficients with dissolved oxygen. Features selected for final modeling ($|r| > 0.4$) are indicated.

Feature Name	Category	Description	Correlation	Selected
season*	Temporal	Hemisphere-aware seasonal classification	N/A	✓
day_of_year	Temporal	Day of year (1-365)	-0.178	✗
sst_anomaly	Temperature	SST deviation from seasonal mean	-0.645	✓
sst_squared	Temperature	SST^2 (non-linear temperature effect)	-0.650	✓
sst_cubed	Temperature	SST^3 (non-linear temperature effect)	-0.606	✗*
log_chlor_a	Log Transform	Log-transformed chlorophyll-a	0.494	✓
log_poc	Log Transform	Log-transformed poc	0.465	✗*
log_Rrs_531	Log Transform	Log-transformed Rrs_531	-0.039	✗
log_Rrs_547	Log Transform	Log-transformed Rrs_547	0.112	✗
log_Rrs_555	Log Transform	Log-transformed Rrs_555	0.161	✗
log_Rrs_645	Log Transform	Log-transformed Rrs_645	0.157	✗
log_Rrs_667	Log Transform	Log-transformed Rrs_667	0.048	✗
log_Rrs_678	Log Transform	Log-transformed Rrs_678	0.034	✗
poc_chla_ratio	Biomass	POC/Chlor_a ratio (biomass efficiency)	-0.361	✗
chlor_poc_ratio	Biomass	Chlor_a/POC ratio (inverse)	0.283	✗
ratio_443_547	Spectral Ratio	Rrs_443/Rrs_547 (blue-green ratio)	-0.474	✓
ratio_443_555	Spectral Ratio	Rrs_443/Rrs_555 (blue-green ratio)	-0.471	✗*
ratio_667_443	Spectral Ratio	Rrs_667/Rrs_443 (red-blue ratio)	0.001	✗
ratio_667_555	Spectral Ratio	Rrs_667/Rrs_555 (red-green ratio)	0.222	✗
ratio_555_667	Spectral Ratio	Rrs_555/Rrs_667 (green-red ratio)	0.005	✗
ndci	Normalized Index	Normalized Difference Chlorophyll Index	-0.010	✗

ndvi_water	Normalized Index	Water-adapted Normalized Difference Vegetation Index	-0.072	✗
sst_chlor_interaction	Interaction	SST x log(chlor_a)	0.467	✓
sst_poc_interaction	Interaction	SST x POC	0.110	✗
abs_latitude	Spatial	Absolute latitude (biogeographic gradient)	0.636	✓

Table D1: Feature Engineering Candidates and Correlation-Based Selection

*: Season is a categorical feature, so correlation was not calculated, but it was retained to capture seasonal patterns in oxygen.
✗*: Initially selected based on correlation threshold but removed due to multicollinearity.

Figure D1 visualizes the correlation magnitudes of all candidate features with dissolved oxygen, illustrating the dominance of temperature-related features and the applied selection threshold.

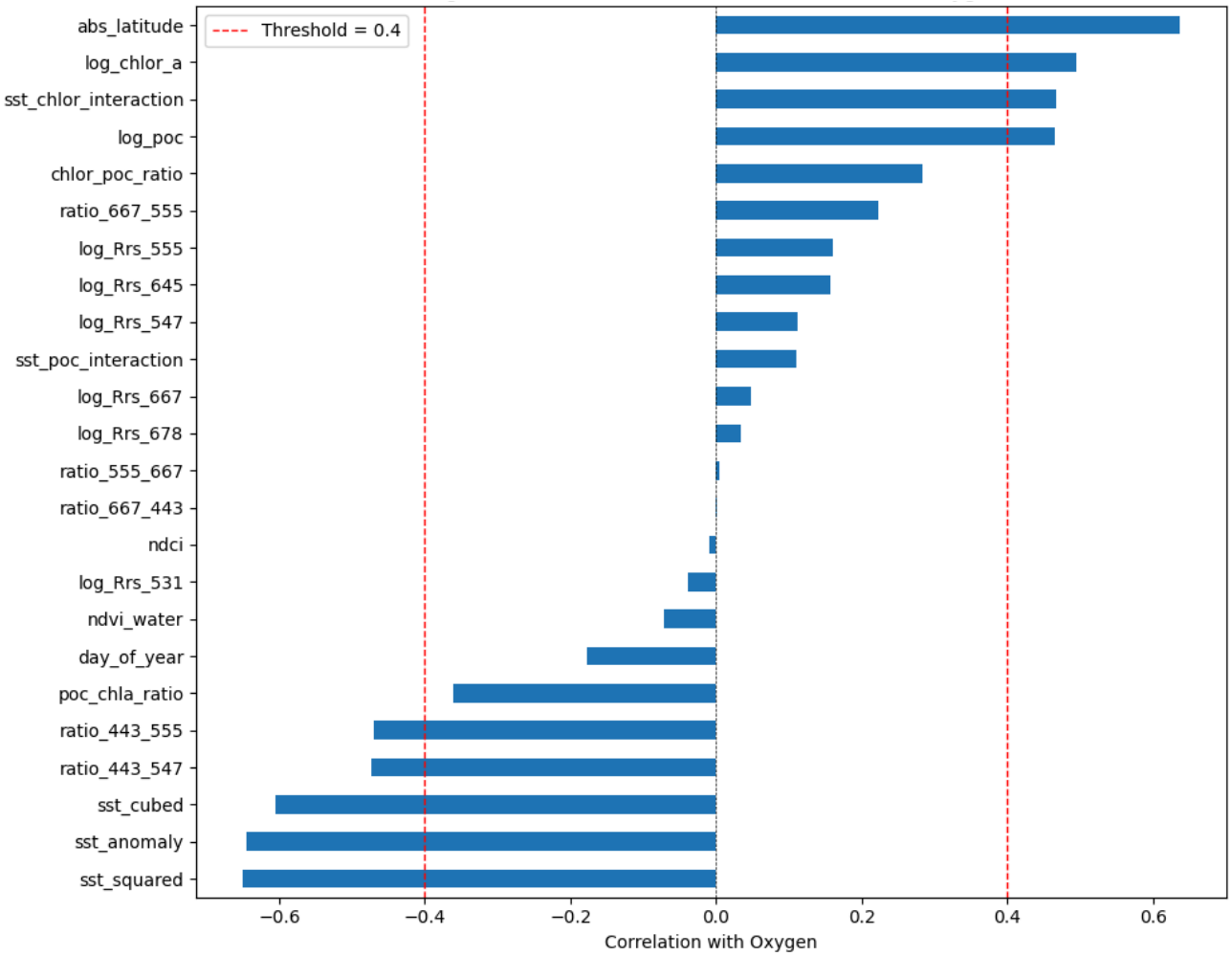


Figure D1: Pearson correlation coefficients between candidate engineered features and dissolved oxygen. Red dashed lines indicates the selection threshold (|r| > 0.4).

Appendix E

The following appendix details the hyperparameter tuning results for all baseline models, achieved using RandomizedSearchCV with TimeSeriesSplit cross-validation. For each model, the search space, parameter function, and selected optimal values are presented.

Elastic Net

Parameter	Meaning	Options	Optimal
L1 Ratio	Ratio of L1 to the combined L1 and L2 penalties	0.1, 0.3, 0.5, 0.7, 0.9	0.3
Alpha	Impact of total penalty	0.001, 0.01, 0.1, 1.0, 10.0	0.01

Table E1: Elastic Net Hyperparameter Configuration

Random Forest

Parameter	Meaning	Options	Optimal
Number of Estimators	Number of trees in the forest	50,100,150	100
Minimum Sample Split	Smallest number of data points needed to split	2, 5	5
Minimum Sample Leaf	Minium number of samples each leaf most have	1, 2	2
Max depth	Maximum a depth a tree is allowed to be	None, 10, 20	20
Bootstrap	Rather or bootstrapping (random sampling with replacement) is used	True	True

Table E2: Random Forest Hyperparameter Configuration

LightGBM

Parameter	Meaning	Options	Optimal
Subsample	Portion of data sampled by each tree	0.7, 0.8, 0.9	0.7
Number of Leaves	Number of leaves by tree	20, 31, 40	40
Number of Estimators	Number of trees in the forest	100, 200, 500	500
Max depth	Maximum depth a tree is allowed to be	5, 7, -1	-1
Learning Rate	Rate at which the model learns	0.01, 0.05, 0.1	0.01
Columns Sampled by Tree	What portions of columns each tree can sample from	0.7, 0.8, 0.9	0.7

Table E3: LightGBM Hyperparameter Configuration

XGBoost

Parameter	Meaning	Options	Optimal
Subsample	Portion of data sampled by each tree	0.7, 0.9	0.7
Reg Lambda	L2 regularization term on weights	1, 1.5, 2	2
Reg Alpha	L1 regularization term on weights	0, 0.005, 0.01	0.005
Number of Estimators	Number of trees in the forest	100, 200, 300	100
Max depth	Maximum depth a tree is allowed to be	3, 5, 7	5
Learning Rate	Rate at which the model learns	0.01, 0.1, 0.2	0.01
Gamma	Minimum loss reduction to make a split	0, 0.1, 0.2	0.1
Columns Sampled by Tree	What portions of columns each tree can sample from	0.7, 0.9	0.9

Table E4: XGBoost Hyperparameter Configuration

Neural Network

Parameter	Meaning	Options	Optimal
Regressor Solver	What solver the model will use	adam, sgd	adam
Max iterations	Maximum number of epochs	200, 500	500
Learning Rate Initialization	The initial setting for the learning rate	0.001, 0.01	0.001
Hidden layer sizes	Size of each layer of the multilayer preceptron	(50,), (100, 50), (200, 100, 50)	(50,)
Alpha	Strength of the L2 regularization	0.0001, 0.001, 0.01	0.0001
Activation	Activation function used by the multilayer preceptron	relu, tanh, logistic	logistic

Table E1: Neural Network Hyperparameter Configuration

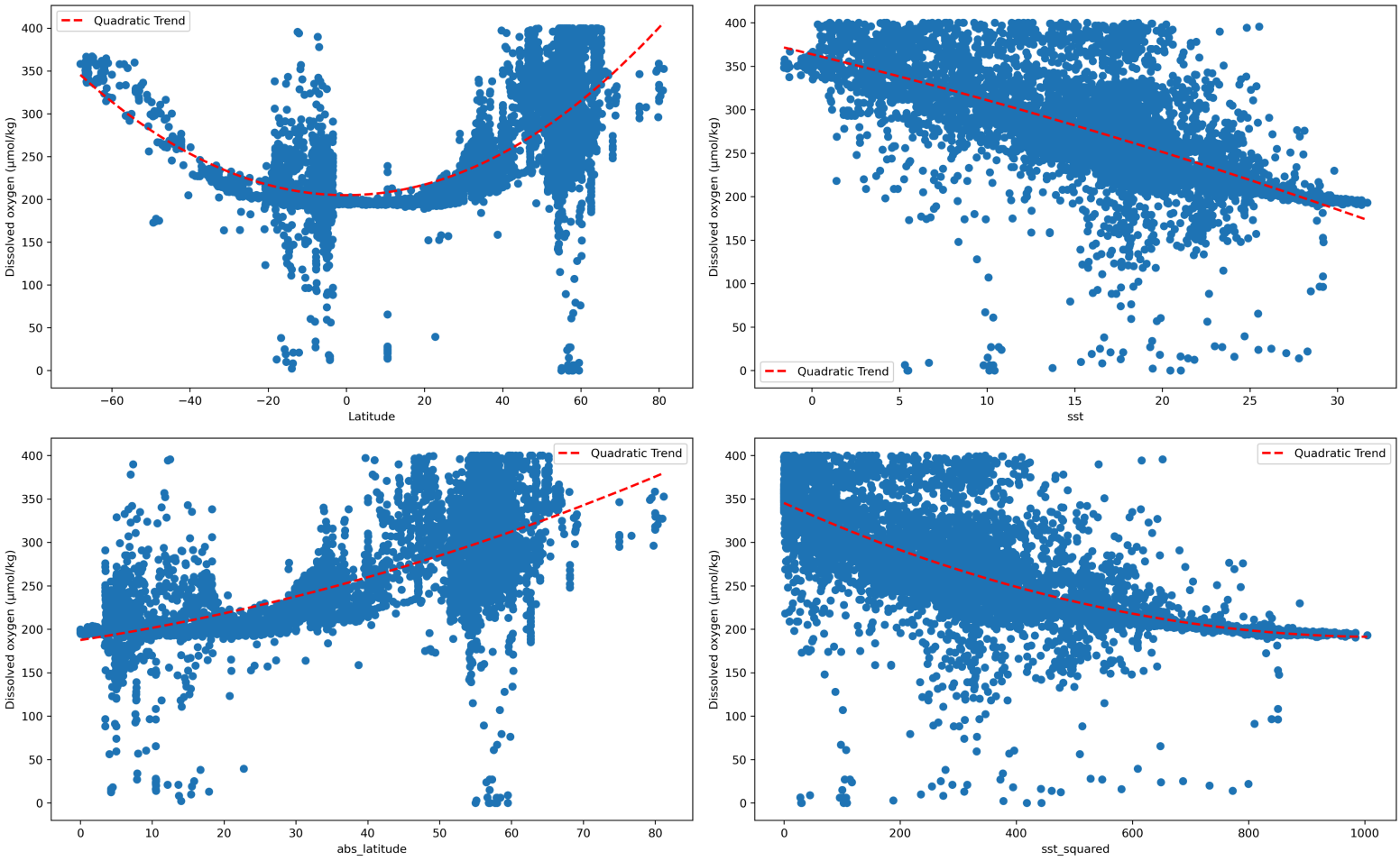
Appendix F

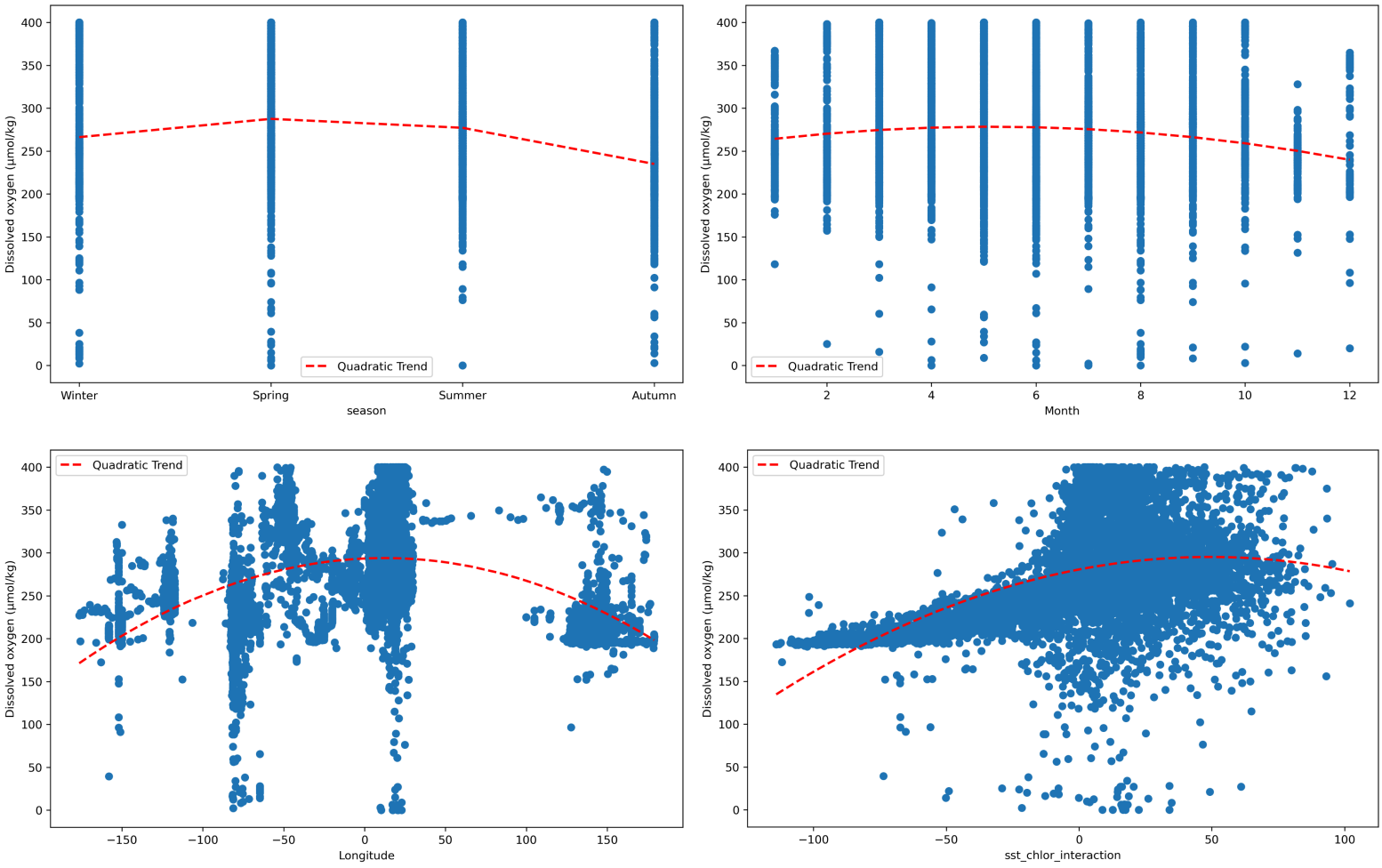
The following appendix contains the comprehensive diagnostic analyses conducted on the final XGBoost model, including feature importance rankings, trend analyses examining relationships between top predictive features and dissolved oxygen concentrations, and an ablation study quantifying the impact of sequential feature removal.

Feature Importance

Feature importance analysis (Figure 3) reveals dominance of spatial (36%), temperature (20%), and temporal (16%) features, with absolute latitude, latitude, and SST ranking as the top three predictors. Satellite spectral measurements contribute minimally (< 2% each), suggesting limited added value from ocean color data beyond what spatial-temporal proxies already capture.

Trend Analysis





Figures F2 - F9: Trend Analysis of Top Features vs. Dissolved Oxygen

Trend analyses validate the model's reliance on spatial and temperature features. Both latitude plots (F2, F4) show clear positive relationships with dissolved oxygen, with concentrations increasing toward the poles. Temperature features (F3 & F5) exhibit strong negative correlations - higher SST corresponds to lower oxygen, consistent with fundamental gas solubility principles. Temporal features (season, month - F6, F7) exhibit weak patterns with substantial overlap across categories. Longitude (F8) and SST-chlorophyll interaction (F9) display weak, noisy relationships. These patterns confirm that spatial and thermal variables drive model predictions, while biological and temporal features contribute minimally.

Ablation Study

An ablation study was conducted to assess the impact of individual features on model performance by sequentially removing the top 10 features in order of importance and retraining the model.

# Removed	Features Removed	R^2	MAE	RMSE
1	['abs_latitude']	0.560	25.853	41.572
2	['abs_latitude', 'sst_squared']	0.556	24.995	41.749
3	['abs_latitude', 'sst_squared', 'sst']	0.545	26.515	42.253
4	['abs_latitude', 'sst_squared', 'sst', 'Month']	0.556	26.300	41.732
5	['abs_latitude', 'sst_squared', 'sst', 'Month', 'Latitude']	0.525	27.356	43.184
6	['abs_latitude', 'sst_squared', 'sst', 'Month', 'Latitude', 'season_Autumn']	0.526	26.494	43.148
7	['abs_latitude', 'sst_squared', 'sst', 'Month', 'Latitude', 'season_Autumn', 'Longitude']	0.490	27.759	44.755
8	['abs_latitude', 'sst_squared', 'sst', 'Month', 'Latitude', 'season_Autumn', 'Longitude', 'season_Winter']	0.500	27.800	44.480
9	['abs_latitude', 'sst_squared', 'sst', 'Month', 'Latitude', 'season_Autumn', 'Longitude', 'season_Winter', 'season_Spring']	0.476	28.501	45.354
10	['abs_latitude', 'sst_squared', 'sst', 'Month', 'Latitude', 'season_Autumn', 'Longitude', 'season_Winter', 'season_Spring', 'sst_chlor_interaction']	0.462	29.487	45.965

Table G1: Findings of Ablation Study on XGBoost Model

Sequential removal of the top 10 features showed R^2 declining from baseline 0.592 to 0.462, a total decrease of 0.130. The largest drops occurred when absolute latitude (step 1: $\Delta R^2 = -0.032$) and longitude (step 7: $\Delta R^2 = -0.036$) were removed, confirming that spatial features are critical predictors. Interestingly, R^2 occasionally increased when month (step 4) and season_Winter (step 8) were removed, suggesting these features either worked better in conjunction with already-removed variables, that other features could substitute for them, or that the model had overfit to these features. The fact that most feature removals decreased R^2 confirms these are strong predictors in their own right, though the modest total performance drop (0.130) indicates model capacity is constrained more by fundamental limitations of the feature set than by any single predictor.

Appendix G

The following appendix evaluates the final XGBoost model's ability to detect hypoxic conditions (DO < 60 µmol/kg) in the held-out test set.

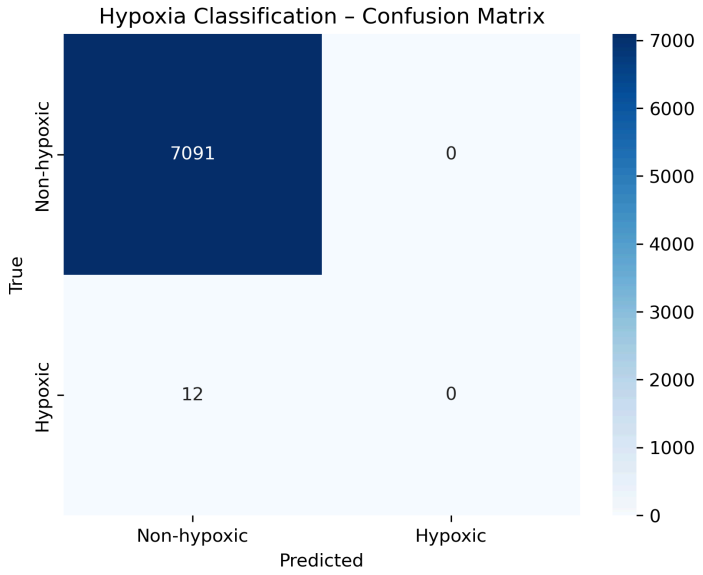
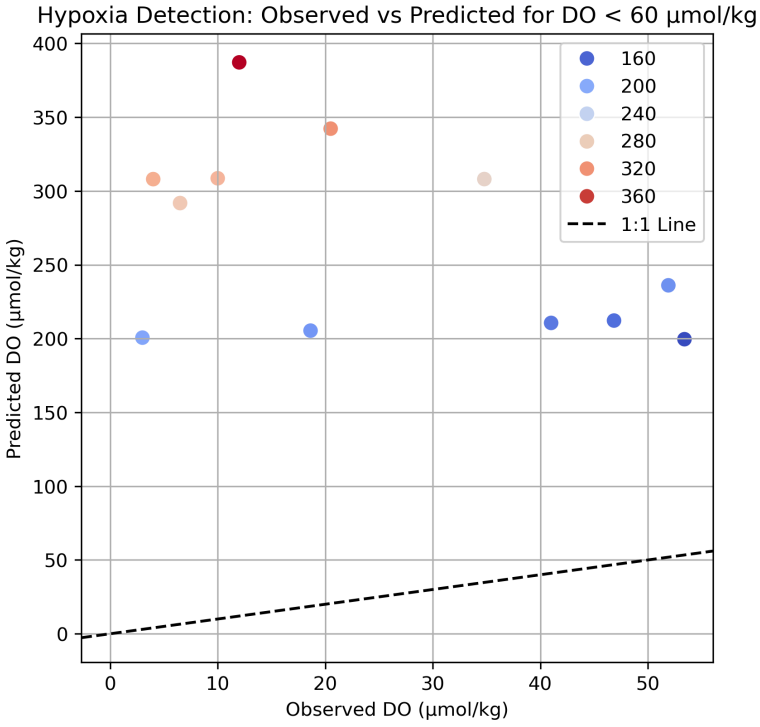


Figure G1: Confusion Matrix for Hypoxia Classification



Figures G2: Observed vs. Predicted DO Values for Samples with Observed DO < 60 µmol/kg.

Both figures above show the performance of the final XGBoost model on hypoxic samples of observed DO < 60 μmol/kg. Figure G1 demonstrates that the model consistently predicts concentrations of 180–350 μmol/kg for hypoxic samples regardless of true values. Of the 12 truly hypoxic samples in the test set, none were correctly identified, with all misclassified as non-hypoxic (Figure G2). This complete detection failure reflects both extreme class imbalance (0.17% hypoxic samples) and the model's inability to predict low oxygen concentrations from surface observations.