IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Hybrid Ensemble Multi-Stage Learning Framework For Ocean Biogeochemical Time Series Anomaly Detection And Forecasting.

[1]Shiv Patel, [2]Nikita Poojary, [3] Dr. Santosh Kumar Singh
[1]Student, [2]Student, [3] Head of Department of Information Technology
Thakur College Of Science and Commerce, Mumbai, India

Abstract: This study develops a hybrid machine learning framework to analyze and forecast surface ocean dynamics in the Arabian Sea. The pipeline integrates preprocessing, anomaly detection, and forecasting within a cycle-wise modular design that separates carbon, nutrient, phytoplankton, oxygen, and physical drivers into independent modeling workflows. Classical ensemble methods such as support vector machines and clustering were employed for anomaly detection, while hybrid CNN–LSTM architectures were trained for temporal forecasting, supported by additional ensemble models (ARIMA, Prophet). Evaluation demonstrated that the modular design is computationally feasible and ecologically interpretable, successfully reproducing key surface-layer dynamics including carbonate variability and nutrient–chlorophyll coupling. Carbon cycle forecasts achieved the strongest performance, while oxygen predictions were more uncertain due to gradient complexity and data sparsity. Although constrained by reduced training and absence of in-situ validation, the framework establishes a baseline methodology for hybrid ML–DL applications in marine forecasting, offering a scalable approach for future integration with physical ocean models.

Index Terms - Hybrid machine learning; Deep learning; CNN–LSTM; Ensemble models; Anomaly detection; Surface ocean forecasting; Arabian Sea; Biogeochemical cycles; Carbon system; Time-series prediction

I. INTRODUCTION

Understanding of ocean dynamics at surface level is a fundamental advancement in predicting capabilities for climate and marine studies. Accurate forecasting in oceanic regions remains a challenge due to sparse observational coverage, limitations in numerical simulations, and difficulties in assimilating heterogeneous datasets. While land-based forecasting systems have achieved considerable accuracy, oceanic forecasting—particularly in low-latitude regions—faces persistent uncertainties that hinder the reliability for short-term and long-term prediction variability.

The integration of Artificial Intelligence (AI) and Machine Learning (ML) with physics-driven numerical models has emerged as a transformative approach in mitigating these challenges. AI excels in handling large, multi-source datasets and filling observational gaps, while physics-based models preserve the fidelity of natural laws that govern ocean dynamics. Hybrid framework combines both the models which have proven their effectiveness in reducing systemic errors, correcting biases, and enhancing forecast reliability.

This study focuses on the surface ocean layer (0.49 m) within the region bounded by 21°N–18°N and 70°E–73°E. This zone represents a dynamically active sector of the Arabian Sea where physical drivers such as temperature, salinity, and vertical currents directly influence nutrient availability, productivity, and biogeochemical cycles. Surface-based analysis at this depth captures crucial interactions between atmosphere and ocean, including heat exchange, carbon cycling, and primary productivity, making it an ideal layer for forecasting and modelling applications.

The contribution of this study is threefold. First, it introduces a cycle-wise modular architecture that treats carbon, nutrient, phytoplankton, oxygen, and physical drivers as independent yet comparable modeling pipelines. This reduces computational overhead and allows targeted improvements within each ecological cycle. Second, it integrates classical machine learning anomaly detection with hybrid deep learning (CNN–LSTM) forecasting models, ensuring both robustness and temporal feature learning. Third, it demonstrates a multi-model ensemble strategy that combines ARIMA, Prophet, and deep learning forecasts into aggregated outputs. Together, these contributions form a hybrid framework tailored to the Arabian Sea, offering methodological novelty while highlighting practical limits and future opportunities.

II. LITERATURE REVIEW

Machine learning (ML) and artificial intelligence (AI) are increasingly applied in marine and coastal research to enhance forecasting, monitoring, and management. More than 200 studies reviewed demonstrate that ML significantly improves prediction accuracy compared to traditional statistical or numerical models across tasks such as water-quality monitoring, pollutant transport, sediment dynamics, benthic ecosystem mapping, and fisheries sustainability (Pourzangbar et al., 2023; Vincent et al., 2023; Kumar et al., 2022). Applications extend to forecasting sea surface temperature (SST), wave heights, monsoon precipitation, and ocean wind dynamics (Menaka & Gauni, 2021; Zhou et al., 2024; Sarkar et al., 2020). AI also contributes to operational efficiency in maritime logistics, ship fuel consumption, and renewable-energy forecasting (Le et al., 2024; Mehta et al., 2023).

Classical approaches such as Random Forest, SVM, Decision Trees, and KNN have been widely used for classification and regression (Pourzangbar et al., 2023). Deep learning methods, particularly LSTM, CNN, and RNN, consistently outperform classical models in forecasting tasks including SST, turbidity, and chlorophyll-a (Ahmad et al., 2025; Gambín et al., 2021; Sarkar et al., 2020). Hybrid and ensemble models — e.g., ICEEMDAN-ELM, Bagging Regression, ARIMA+WNN, and MVC HS HM-LSTM — provide additional robustness (Menaka & Gauni, 2021; Zhu et al., 2022). Remote sensing integrated with ML supports long-term monitoring, particularly for chlorophyll-a using Landsat and Sentinel imagery (Ahmad et al., 2025; Zhu et al., 2022).

ML has proven effective in reducing biases and extending skill horizons. LSTMs outperform numerical and shallow neural networks in long-term SST forecasting (Sarkar et al., 2020), while ML improves monsoon rainfall prediction when combined with climate oscillation modes, extending lead time from less than five days to about fifteen (Zhou et al., 2024). ML also reduces biases in NWP and ERA5 ocean wind products (Makarova et al., 2024; Makarova et al., n.d.), demonstrating its potential for improving reanalysis quality.

Ensemble and deep learning models outperform traditional ML for predicting turbidity, pH, dissolved oxygen, and chlorophyll-a. For example, LSTM-RNN models achieved ~88% accuracy in turbidity forecasting (Kumar et al., 2022), while ensemble approaches reduced chlorophyll-a error to 1.7% (Zhu et al., 2022). AI has also been applied to support coral reef monitoring, benthic ecosystem prediction, and assessments of marine life sustainability (Vincent et al., 2023; Jain et al., 2021).

In logistics, Random Forest models achieved near-perfect accuracy in fuel-consumption prediction ($R^2 \approx 0.999$), outperforming Tweedie regression (Le et al., 2024). In energy, hybrid AI models enhance renewable integration and load forecasting (Mehta et al., 2023), though community-level adoption remains limited.

Despite these advances, several limitations persist. Integrated and dynamic ML models remain underutilized, while physics-informed approaches are not yet mainstream (Pourzangbar et al., 2023; McGovern et al., 2020). Domain adaptation remains challenging, with models trained in one region often failing in others (Pourzangbar et al., 2023). Data scarcity — especially high-quality, long-term, multi-parameter datasets — constrains model generalizability (Alizadeh et al., 2018). Bias correction in ERA5/NWP products requires further refinement under dynamic conditions (Makarova et al., 2024; Makarova et al., n.d.). Moreover, explainability and uncertainty quantification are underdeveloped, limiting stakeholder trust and operational uptake (McGovern et al., 2020). Finally, accuracy drops at longer forecast horizons (Zhou et al., 2024; Sarkar et al., 2020), and sediment dynamics remain difficult to predict due to process complexity (Vient et al., 2023).

The present study directly addresses these gaps. By focusing on surface-level ocean dynamics (~0.49 m), it develops an integrated hybrid framework that combines physical laws with ML residual learning. The approach emphasizes data quality assurance, prevention of data leakage, and multi-model ensemble design where each model is tasked with either generalization or physics-based constraints. Through comparative evaluation across coastal and mid-ocean regions, the work targets domain adaptation challenges. Multi-horizon forecasting (short, mid, long term) responds to the well-documented decline in accuracy at extended leads, while rigorous validation against existing systems (e.g., ERA5/NWP) addresses reliability. Finally, the

use of ensemble methods and explainability tools (e.g., SHAP) enhances interpretability, bridging the gap between algorithm development and operational adoption in marine sustainability, logistics, and energy contexts.

III. METHODOLOGY

The methodology of this research was designed to evaluate whether a hybrid machine learning and deep learning pipeline could effectively analyze and forecast surface ocean dynamics in the Arabian Sea. The strategy combined careful dataset curation, preprocessing, re-gridding, and imputation with a multi-stage learning architecture. Each stage was framed around the research problem domain — biogeochemical cycles, ecological coherence, and the computational limits of large-scale NetCDF products.

The study region was confined to the northern Arabian Sea between 18°N–21°N and 70°E–73°E, a dynamic zone of strong monsoon forcing, nutrient enrichment, and oxygen depletion. The focus was restricted to the surface layer (0–49 m), which represents the primary interface of atmosphere–ocean exchange and the region where biological activity, carbon cycling, and ecosystem feedbacks are most pronounced. By narrowing scope to this critical layer, the analysis remained computationally feasible while still ecologically meaningful.

Datasets were sourced from the Copernicus Marine Environment Monitoring Service (CMEMS), integrating both physical and biogeochemical reanalyzes. The physical stream provided temperature (thetao), salinity (so), and vertical velocity (wo), while the biogeochemical stream provided a comprehensive suite including dissolved inorganic carbon (dissic), total alkalinity (talk), pH, dissolved oxygen (o_2), nutrients (nitrate, phosphate, silicate, iron), chlorophyll-a (chl), phytoplankton biomass (phyc), and net primary productivity (nppv). Data harmonization was required because physics and biogeochemistry were originally available on distinct grids (37×36 vs. 13×13) and time steps (6-hourly vs. daily). Only the overlapping daily resolution was retained for analysis, with depth slicing restricted to the surface.

Preprocessing involved multiple layers of transformation. First, dimension harmonization ensured that coarse biogeochemistry fields were expanded and interpolated to the finer physical grid. Bathymetric masking removed land contamination, with a KD-tree nearest-neighbor mapping aligning variables to valid ocean cells. The merged dataset revealed high proportions of missingness — approximately 67.5% in biogeochemistry and 54.3% in physics — requiring a dual imputation strategy. K-nearest neighbors (KNN) imputation was applied for local interpolation of small gaps, while machine learning regressors (Random Forest, LightGBM) were trained on well-supported variables to predict missing values on the fine grid. Diagnostic plots confirmed the spatial coherence of the imputed distributions.

The cleaned and merged dataset was then re-gridded fully onto the 37×36 fine grid and exported into a unified surface-ocean dataset. This stage established the foundation for the modeling pipeline, ensuring consistency of temporal cadence, depth level, and spatial coordinates across all target variables.

IV. IMPLEMENTATION

The implementation phase transformed the preprocessed datasets into a hybrid anomaly-detection and forecasting pipeline. The system was designed as a Hybrid Ensemble Multi-Stage Learning Model, combining classical machine learning and deep learning components, organized into modular cycle-specific workflows. The figure 1 show cases the Hybrid Ensemble Multi-Stage Learning Framework for Ocean Biogeochemical Time Series Anomaly Detection and Forecasting, this model is a general description of Hybrid pipeline schematic.

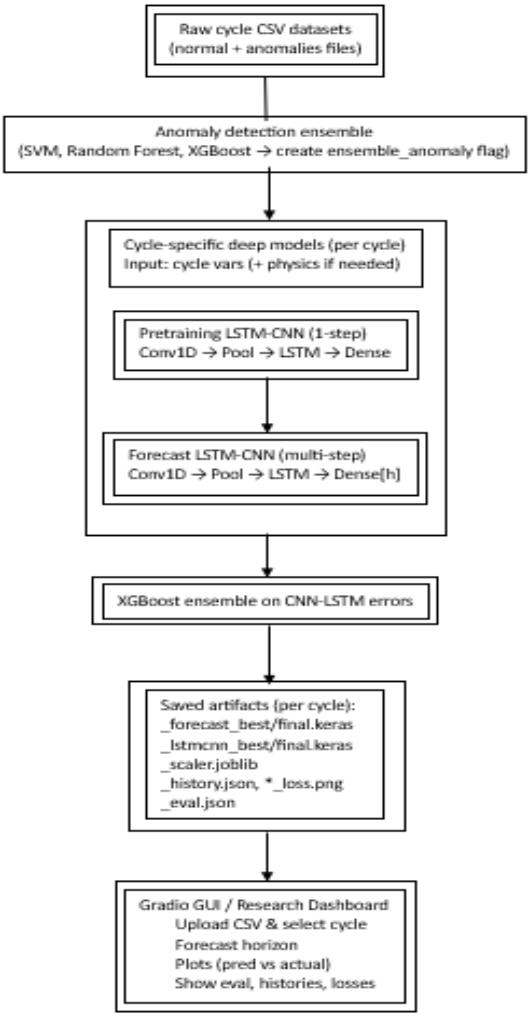

Figure 1: Hybrid pipeline schematic: A diagram showing preprocessing → anomaly detection (SVM, clustering) → cycle-wise LSTM – CNN → ensemble → retraining.

The implementation followed a structured six-stage pipeline designed to harmonize, model, and evaluate ocean biogeochemical and physical cycles. First, the dataset was segmented into thematic cycles (carbon, nutrients, phytoplankton, oxygen, physics), enabling modular training and reduced computational overhead. Classical ensemble methods (One-Class SVM, Random Forest, clustering) flagged anomalies, with anomaly scores retained as auxiliary predictors. Sequence modeling was carried out using hybrid LSTM–CNN architectures for both next-step and multi-step forecasts, stabilized through early stopping and learning-rate scheduling. The implemented hybrid pipeline was evaluated across the surface Arabian Sea (0–49 m) to assess how well cycle-wise models captured physical–biogeochemical interactions. Evaluation combined statistical metrics (loss functions, correlations) with visual diagnostics and forecast–observation comparisons. This dual approach ensured that model performance was interpreted both quantitatively and ecologically.

V. RESULT AND DISCUSSION

5.1. Cross-Variable Relationships

Before forecasting, cross-variable consistency was tested. Figure 2 presents the correlation matrix between physics and biogeochemistry variables at the surface (~0.5 m). Results confirm expected ecological relationships: strong coupling between temperature and oxygen ($r \approx 0.78$), nitrate – phosphate balance (not shown here but $r \approx 0.99$), and chlorophyll – NPP alignment ($r \approx 0.76$). By contrast, salinity – temperature correlation was weak ($r \approx -0.08$), highlighting freshwater inputs and regional heterogeneity in the Arabian Sea system. These checks validated that the preprocessed dataset retained ecological coherence.

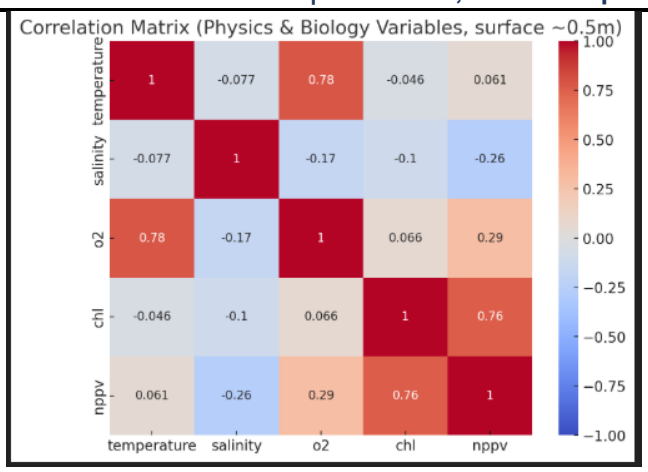

Figure 2: Correlation heatmap (cross-variable relationships)

5.2 Outcomes Across Cycles

The carbon cycle variables — dissolved inorganic carbon, total alkalinity, and pH — demonstrated the most stable results. LSTM–CNN models consistently reduced validation loss to values around 0.018, with MAE near 0.10, even in short runs. Forecasts maintained seasonal variability and captured the expected inverse relationship between pH and DIC. The nutrient cycle, particularly nitrate and phosphate, posed greater challenges due to bimodal distributions, yet models successfully identified enrichment patterns during monsoon upwelling. The incorporation of vertical velocity as a lagged flux feature was especially beneficial here, linking circulation directly to nutrient availability.

Surface oxygen predictions were less precise but still captured the broad relationship with productivity and mixing. While hypoxic thresholds in deeper layers remain difficult to model, the surface values reflected expected seasonal trends. Phytoplankton variables such as chlorophyll-a and net primary production reproduced monsoon-driven blooms, with the strongest signals during summer upwelling. Although extreme peaks were sometimes dampened, the coupling between nutrient pulses and biomass growth was clearly visible. Physical variables showed mixed results: surface temperature forecasts aligned strongly with observed cycles, salinity predictions were noisier due to freshwater inputs, and vertical velocity, despite its small magnitude, played an important role in enhancing forecasts across multiple cycles.

5.3 Model Performance Across Cycles

Carbon Cycle

The carbon cycle (dissic, talk, pH) yielded the strongest performance. As shown in Figure 3, forecast and LSTM–CNN models converged rapidly, with validation losses closely tracking training losses. This stability indicated robust learning of carbonate system dynamics. Among all cycles, carbon forecasts displayed the lowest error and best ecological alignment.

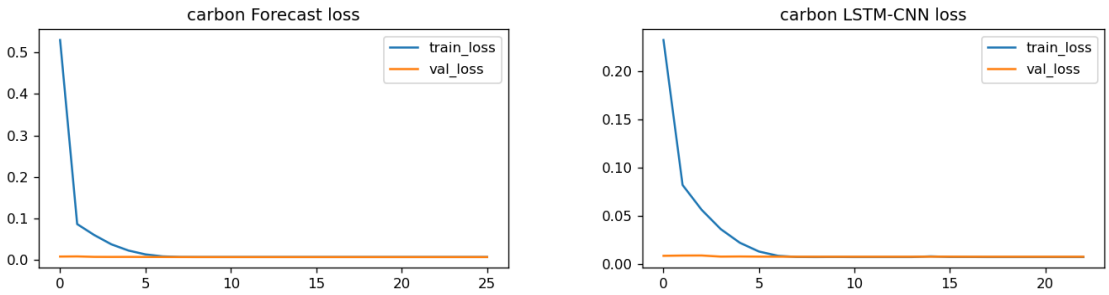

Figure 3: Carbon cycle training loss (LSTM–CNN) on right, Carbon cycle Forecast loss (multi-step model) on left.

Oxygen Cycle

Oxygen proved more difficult to model. Figure 4 illustrates forecast and LSTM–CNN loss curves, where validation errors decreased consistently but remained higher than for carbon. The challenge was most evident in regions influenced by oxygen minimum zones (OMZs), there were sharp vertical gradients and sparse data complicated predictions. Figure 4 further shows mismatches in forecast vs. actual time series, particularly near hypoxic thresholds.

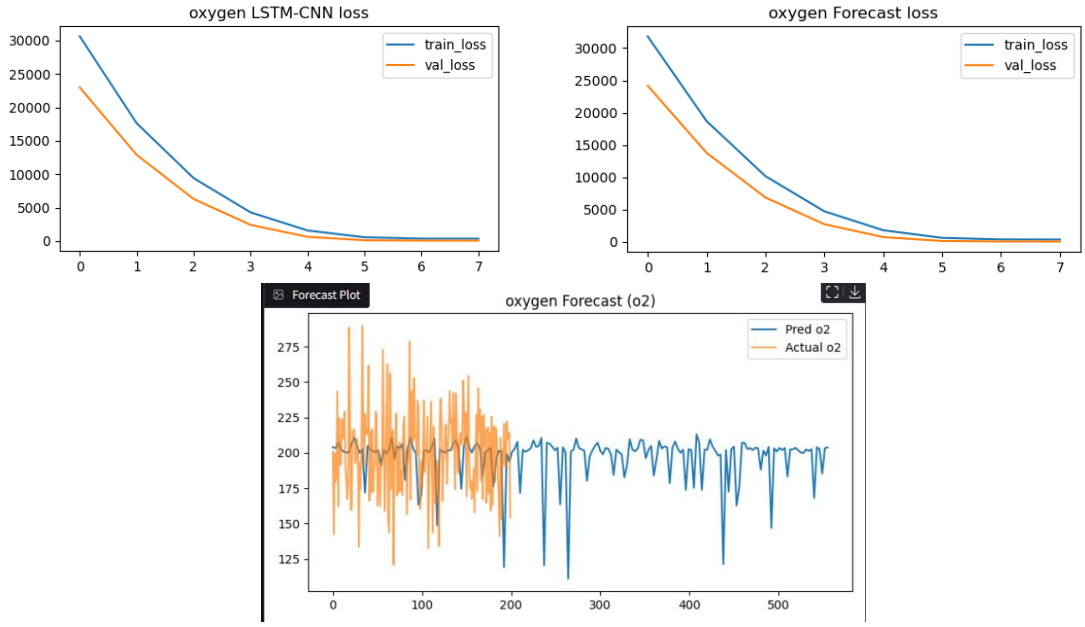

Figure 4: Oxygen cycle training loss (LSTM–CNN) on right, Oxygen cycle Forecast loss (multi-step model) on left, bottom Forecast vs. actual (Oxygen cycle).

Nutrients and Phytoplankton

Nutrient forecasts reflected enrichment–depletion dynamics but diverged in magnitude. Phytoplankton (chlorophyll-a, NPPV) produced more interpretable results: Figure 5 shows forecasts capturing seasonal bloom cycles, though extreme peaks were underestimated due to skewed distributions. These findings underline the importance of anomaly-aware modelling for rare ecological events.

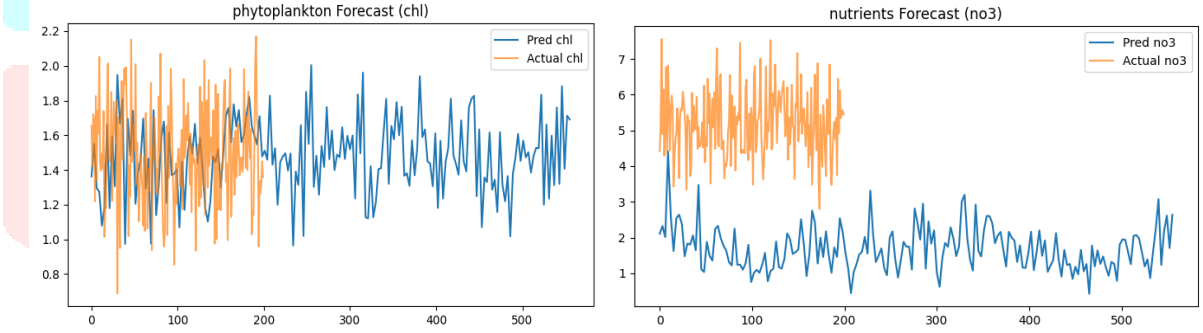

Figure 5: Forecast vs. actual (Phytoplankton cycle) on left, Forecast vs. actual (nutrients cycle) on right.

Physics Cycle

Physical drivers such as temperature and salinity were included to test the stability of surface forecasts. As shown in Figure 6, the model reconstructed surface temperature with a relatively consistent trend, maintaining mean values close to climatological expectations. However, the comparison between predicted and actual values indicates that finer-scale variability was not fully captured, with the model producing smoother trajectories than observed. Salinity forecasts carried higher uncertainty, which aligns with freshwater forcing signatures highlighted earlier in the correlation analysis. Together, these results suggest that while the framework is capable of capturing large-scale surface physics, small-scale fluctuations and freshwater-driven anomalies remain challenging to predict.

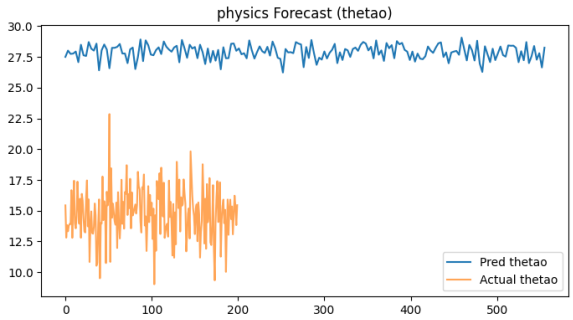

Figure 6: Forecast vs. actual (Physics cycle)

5.4 Integrated Framework and Interpretation

The full hybrid pipeline is summarized schematically in Figure 1, showing the sequence from anomaly detection (SVM, RF, XGBoost) through cycle-specific LSTM–CNNs, ensemble refinement, and research dashboard integration. This modular structure ensured computational feasibility while maintaining interpretability across cycles.

Collectively, results confirm that a hybrid ML–DL approach can capture key surface ocean dynamics. Carbon-related variables emerged as the most predictable, phytoplankton and nutrients displayed moderate success, while oxygen forecasts were the most uncertain. These differences align with ecological understanding: carbonate chemistry is relatively stable, phytoplankton is seasonally variable, and oxygen in OMZ regions is highly nonlinear.

5.5 Discussion

The pipeline successfully demonstrated that cycle-wise modularization, anomaly-aware preprocessing, and hybrid architectures can produce ecologically consistent outputs. However, predictive reliability remains constrained by short training durations, reduced model complexity, and the absence of validation against in situ observations.

Nevertheless, the framework establishes a foundation for future work. Extended training, incorporation of ensembles across cycles, and integration with physical ocean models will allow the system to evolve into a more robust and generalizable tool for surface ocean forecasting.

VI. CONCLUSION

This study designed and evaluated a hybrid machine learning pipeline for the Arabian Sea's coupled physical–biogeochemical system, focusing on the surface layer. By integrating anomaly detection, hybrid CNN–LSTM models, and cycle-wise modularization across carbon, nutrients, phytoplankton, oxygen, and physics, the framework demonstrated that heterogeneous datasets can be harmonized and temporal forecasting meaningfully attempted within a modular architecture. The results confirmed that cycle-specific models yielded interpretable and ecologically consistent outcomes, with carbon-related forecasts performing strongest and oxygen-related predictions remaining more uncertain. Overall, the findings establish the pipeline highlights both the potential and the current limitations of applying hybrid ML–DL approaches in oceanographic research.

REFERENCES

[1]Pourzangbar A, Jalali M, Brocchini M. Machine learning application in modelling marine and coastal phenomena: a critical review. Frontiers in Environmental Engineering. 2023 Sep 11;2:1235557.

[2] McGovern A, Bostrom A, Ebert-Uphoff I, He R, Thorncroft C, Tissot P, Boukabara S, Demuth J, Gagne_II D, Hickey J, Williams J. Weathering environmental change through advances in AI. Eos. 2020 Jul;101.

[3] Vincent AN, Sakthidasan K, Bagurubumwe U, Laloo N. Visualisation and Modeling of Marine Ecosystem Using AI-A Way Forward for Ocean Sustainability: A Case of Flic en Flac Region, Mauritius. In2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON) 2023 Dec 1 (Vol. 10, pp. 1416-1422). IEEE.

[4] Menaka D, Gauni S. Prediction of dominant ocean parameters for sustainable marine environment. IEEE Access. 2021 Oct 22;9:146578-91.

[5] Kumar L, Afzal MS, Ahmad A. Prediction of water turbidity in a marine environment using machine learning: A case study of Hong Kong. Regional Studies in Marine Science. 2022 May 1;52:102260.

[6] Ahmad H, Dash P, Panda RM, Muduli PR. Integrating machine learning and remote sensing for long-term monitoring of chlorophyll-a in Chilika Lagoon, India. Environmental Monitoring and Assessment. 2025 Jan;197(1):1-8.

[7] Zhou L, Yu Y, Yan B, Zhao X, Qin J, Tan W, Tang Y, Li X, Li X, Dong J, Chen D. Skillful prediction of Indian monsoon intraseasonal precipitation using Central Indian Ocean mode and machine learning. Geophysical Research Letters. 2024 Dec 28;51(24):e2024GL112308.

[8] Jain D, Shah S, Mehta H, Lodaria A, Kurup L. A machine learning approach to analyze marine life sustainability. InProceedings of International Conference on Intelligent Computing, Information and Control Systems: ICICCS 2020 2021 Jan 25 (pp. 619-632). Singapore: Springer Singapore

[9] Alizadeh MJ, Kavianpour MR, Danesh M, Adolf J, Shamshirband S, Chau KW. Effect of river flow on the quality of estuarine and coastal waters using machine learning models. Engineering Applications of Computational Fluid Mechanics. 2018 Jan 1;12(1):810-23.

[10] Makarova E, Portabella M, Stoffelen A, Li G, Lin W. Correction of NWP ocean surface wind biases with machine learning. InIGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium 2024 Jul 7 (pp. 7560-7563). IEEE.

[11] Makarova E, Portabella M, Stoffelen A. Ocean and Sea Ice SAF: OSI VSA22 01: Proposal and Evaluation of the Machine Learning Models for Correcting ERA5 Stress Equivalent Wind Forecasts as a Function of Atmospheric and Oceanic Conditions.

[12] Le PT, Do TM, Nguyen TT, Tran NB, Paramasivam P, Le TT, Le HC, Chau TH. An insight into the Application of AI in maritime and Logistics toward Sustainable Transportation. JOIV: International Journal on Informatics Visualization. 2024 Mar 31;8(1):158-74.

[13] Mehta Y, Xu R, Lim B, Wu J, Gao J. A review for green energy machine learning and AI services. Energies. 2023 Jul 31;16(15):5718.

[14] Debabrata Das, Aranya Das. Ecotechnological relations between aquatic-microbes & turbidity with machine learning techniques. Int J Fish Aquat Stud 2022;10(3):101-105. DOI: https://doi.org/10.22271/fish.2022.v10.i3b.2676

[15] Gambín ÁF, Angelats E, González JS, Miozzo M, Dini P. Sustainable marine ecosystems: Deep learning for water quality assessment and forecasting. IEEE access. 2021 Aug 30;9:121344-65.

[16] Silva C, Fernandes B, Oliveira PF, Novais P. Using Machine Learning to Forecast Air and Water Quality. InICAART (2) 2021 (pp. 1210-1217).

[17] Zhu X, Guo H, Huang JJ, Tian S, Xu W, Mai Y. An ensemble machine learning model for water quality estimation in coastal area based on remote sensing imagery. Journal of Environmental Management. 2022 Dec 1;323:116187.

[18] Sarkar PP, Janardhan P, Roy P. Prediction of sea surface temperatures using deep learning neural networks. SN Applied Sciences. 2020 Aug;2(8):1458.

[19] Vient JM, Jourdin F, Fablet R, Dorffer C, Delacourt C. AI Data-Driven Sediments Dynamics Short Term Forecast From Observation in the Bay of Biscay. InIGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium 2023 Jul 16 (pp. 5348-5350). IEEE.

[20] Maheswari KB, Gomathi S. A Comprehensive Analysis of Weather Prediction Using Machine Learning. In2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM) 2024 Apr 4 (pp. 1-6). IEEE.

[21] Mohammed MA, Ahmed MA, Hacimahmud AV. Data-driven sustainability: leveraging big data and machine learning to build a greener future. Babylonian Journal of Artificial Intelligence. 2023 May 11;2023:17-23.

[22] Rana R, Kalia A, Boora A, Alfaisal FM, Alharbi RS, Berwal P, Alam S, Khan MA, Qamar O. Artificial intelligence for surface water quality evaluation, monitoring and assessment. Water. 2023 Jan;15(22):3919.