Archives available at journals.mriindia.com

International Journal on Mechanical Engineering and Robotics

ISSN: 2321-5747

Volume 14 Issue 02, 2025

Integration of Satellite Remote Sensing and Machine Learning for Pond Water Quality Prediction in Eluru

[1]Dr. Aparitosh Gahankari, [2]Komal Waghade, [3]Aayushi Joshi, [4]Kashish Taklikar, [5]Aachal Raut

[1,2,3,4,5] *Department of Artificial Intelligence, St. Vincent Pallotti College of Engineering and Technology, Nagpur*

Peer Review Information

Submission: 05 Nov 2025

Revision: 25 Nov 2025

Acceptance: 17 Dec 2025

Keywords

Remote Sensing, Sentinel-2, Water Quality Prediction, Aquaculture, Machine Learning, Random Forest, Spectral Indices

Abstract

Aquaculture is essential for addressing global food security, but its sustainable expansion faces significant hurdles—particularly in monitoring water quality. In key aquaculture hubs like Eluru, India, the well-being and output of fishponds hinge on critical factors like dissolved oxygen (DO), pH levels, and ammonia concentrations. Conventional monitoring techniques, which involve labor-intensive manual sampling, are costly, inefficient, and difficult to scale across vast pond networks. To overcome these limitations, this study introduces an innovative solution: a fusion of satellite remote sensing and machine learning designed to deliver scalable, affordable, and near-instantaneous water quality assessments.

At the core of this approach is Sentinel-2 multispectral imagery, offering detailed optical data spanning visible, near-infrared (NIR), and shortwave infrared (SWIR) wavelengths. While DO, ammonia, and pH cannot be directly measured by satellites, they can be estimated using spectral indicators such as reflectance values (B2, B3, B4, B8, B11, B12) and water-vegetation indices like NDVI, NDWI, MNDWI, and NDCI. The methodology follows a two-phase process: (i) reconstructing missing satellite data caused by cloud cover or gaps using interpolation and regression techniques to maintain dataset continuity; and (ii) training a Random Forest regression model on historical in-situ measurements alongside satellite-derived metrics to predict DO, ammonia, and pH concurrently.

Findings reveal that this combined method effectively compensates for incomplete satellite observations while yielding precise estimates of vital water quality metrics. By facilitating large-scale, routine monitoring of thousands of ponds, the system drastically reduces reliance on manual sampling. This advancement holds promise for bolstering aquaculture welfare programs, enhancing fish health and productivity, and promoting the long-term viability of inland aquaculture operations.

Introduction

The production of food globally is increasingly relying on the aquatic sector, and fish farming (aquaculture) has become one of the most rapidly expanding contributors, holding a crucial position in maintaining the availability of nutritious food supplies worldwide. In India, this activity is vital for local economies, supporting rural employment, driving exports, and fostering general economic development. The success of these systems, particularly those relying on inland ponds, is intrinsically dependent on

© 2025 The Authors. Published by MRI INDIA.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

diligent water condition management. Key physicochemical factors, such as the concentration of dissolved oxygen (DO), the pH balance, and the level of ammonia, directly govern the health, growth trajectories, and ecological stability of the fish population. Keeping these indicators stabilized within narrow limits is imperative; deviation often triggers stress responses, elevates susceptibility to disease, and can ultimately result in catastrophic mass mortality events.

Historically, the monitoring of water parameters has necessitated conventional physical sampling paired with on-site measurements. While adequate for limited operations, this technique presents severe constraints. Physical testing is overly reliant on labor, consumes extensive time, and lacks the necessary scope when applied to the millions of active aquaculture sites distributed across states like Andhra Pradesh. Furthermore, the sporadic nature of these evaluations—frequently performed on a weekly or monthly basis—renders them ineffective at identifying rapid, critical shifts in water composition. Consequently, essential remedial actions are often delayed, compromising both farm output and the well-being of the stock. Overcoming this difficulty requires a strategic shift toward monitoring solutions that are scalable, cost-effective, and capable of providing data in near real-time.

Satellite-based Earth observation technology presents a compelling methodology for large-scale, non-destructive surveillance of aquatic habitats. The Sentinel-2 mission, a multi-band satellite system, delivers high spatial resolution imagery with a recurring observational cycle, making it particularly well-suited for tracking continental water masses. Although parameters such as DO, ammonia, and pH lack optical interaction and cannot be gauged directly from orbit, they can be accurately deduced via correlated indicators, or proxies. Analysis of spectral signatures in specific wavelength channels (e.g., the Blue, Green, Red, Near-Infrared, and Shortwave Infrared bands) and calculated metrics (including the Normalized Difference Vegetation Index (NDVI), the Normalized Difference Water Index (NDWI), the Modified NDWI (MNDWI), and the Normalized Difference Chlorophyll Index (NDCI)) provides essential information on phytoplankton density, chlorophyll content, and the water's clarity. These ecological factors exhibit strong correlation with the targeted water quality metrics.

This project outlines a synergistic methodology that fuses remote sensing capabilities with computational intelligence to forecast critical water quality variables within aquaculture ponds. This approach is structured into two main phases. First, we tackle the persistent issue of data scarcity arising from cloud cover or unavailability of scenes on specific dates. This involves utilizing sophisticated imputation and regression models to estimate missing spectral values, thereby guaranteeing an unbroken data stream for subsequent model training. Second, we employ a Random Forest regression algorithm, trained iteratively on historical, localized ground observations of DO, ammonia, and pH alongside the derived satellite features (spectral bands and indices). This system is engineered for the synchronous prediction of multiple water quality parameters across any specified time and location.

The resulting monitoring architecture dramatically enhances operational scalability, facilitating oversight of thousands of ponds with minimal manpower. Unlike manual surveillance, which faces severe restrictions from time and labor costs, the satellite-enabled system offers superior temporal fidelity (aligned with Sentinel-2's approximate five-day recurrence) and comprehensive spatial coverage. Furthermore, by integrating machine learning, the system gains the ability to map intricate, non-linear relationships between the spectral proxies and the true water conditions, leading to significantly improved predictive accuracy.

The utility of this research extends beyond routine aquaculture management. By merging Earth observation with predictive modeling, the proposed framework strongly supports environmentally sound farming goals, reduces financial uncertainties for producers, and promotes higher standards of aquatic animal welfare. In geographical areas where physical water quality data acquisition is limited or sporadic, such a tool can function as a vital pre-emptive alert mechanism, enabling prompt intervention before significant ecological degradation occurs. Additionally, the underlying framework is inherently adaptable, allowing it to be applied to other geographical locations or customized for different environmental parameters, thereby offering a highly scalable solution for water quality management across diverse aquaculture settings.

In essence, this study demonstrates the immense potential unleashed by combining orbital monitoring techniques with machine learning to resolve the most pressing challenge confronting aquaculture—efficient, high-volume water quality supervision. This innovative strategy not only compensates for the profound deficiencies of manual methods but also establishes a pathway toward a sophisticated, data-driven

management paradigm capable of ensuring sustainable farming practices and reinforcing global food security both within India and internationally.

Literature Review

The rapid expansion of the aquaculture industry has intensified the need for sophisticated monitoring tools to safeguard water quality and ensure the well-being of farmed fish. Critical parameters influencing pond ecosystems, such as dissolved oxygen (DO), pH, and ammonia, have been consistently identified by research as paramount. Historically, water quality assessments relied predominantly on manual, field-based sampling. While adequate for small-scale operations, these methods are labor-intensive and impractical for widespread deployment across the vast number of aquaculture ponds found in India and other key producing regions. Consequently, academic and industrial efforts have increasingly concentrated on leveraging satellite remote sensing and machine learning as more scalable and efficient alternatives.

1. Remote Sensing for Water Quality Monitoring

Satellite remote sensing has been widely explored for its potential use in water quality assessments. Initial studies successfully highlighted the capabilities of optical sensors for detecting important water quality parameters, which include chlorophyll concentration, suspended sediments, and turbidity. Amongst a host of available platforms, Sentinel-2 has gained considerable popularity due to its fine spatial resolution ranging from 10 to 20 meters and frequent revisits. Furthermore, specific spectral bands, such as B2 (blue), B3 (green), B4 (red), and B8 (near-infrared), have identified relations with chlorophyll's absorption and reflection pattern, while shortwave infrared bands (B11 and B12) are sensitive to suspended particulate matter and turbidity. In addition, some important derived spectral indices, such as NDVI, NDWI, and MNDWI, are widely used to describe plant-water interaction and aquatic biomass changes. The Normalized Difference Chlorophyll Index has become prominent due to its established capability of identifying phytoplankton blooms, which are closely related to the dynamics of dissolved oxygen.

However, direct measurement of optically inactive parameters such as DO, ammonia, and pH through remote sensing remains a challenging task. Instead, these would have to be indirectly inferred by identifying their relationships with optically active proxies like chlorophyll, turbidity, and suspended organic matter. The need for this underlines the importance of advanced statistical and machine learning models able to discern these complex nonlinear relationships.

2. Machine Learning in Aquatic Monitoring

Machine Learning has now become a powerful approach in environmental monitoring, especially in cases where the predictive relationship of variables to outcomes is fundamentally complex and multidimensional. Algorithms like Random Forest, Support Vector Machines, Gradient Boosting, and Neural Networks have seen wide applications in forecasting water quality. The introduction of Random Forest embedded inherent robustness in handling high-dimensional datasets, adding to its resilience to overfitting, established it as one of the favorite algorithms in many remote sensing investigations.

A growing number of studies have combined remote sensing features with machine learning for the prediction of different water quality indicators. For instance, several have illustrated that the ML models, trained on the Landsat imagery, can make fairly accurate predictions of chlorophyll-a concentrations in inland water bodies. In the same way, Sentinel-2 bands coupled with Random Forest regression for turbidity and nutrient level estimation in freshwater lakes. These studies collectively bring out the considerable potential of ML to bridge the analytical gap between satellite-derived spectral data and ground-truth water quality measurements.

Nonetheless, specific research on the integration of satellite data with ML models in aquaculture contexts is still relatively limited. For example, investigation on the relationship between MODIS imagery and productivity in Bangladesh ponds, while others used machine learning methods with features from Sentinel-2 to estimate water quality in aquaculture ponds in India. However, most of these pioneering studies have focused on predicting only one parameter and often without a proper approach to predicting multiple parameters of water quality simultaneously.

3. Addressing Missing Data in Remote Sensing

A significant hurdle in deploying remote sensing for aquaculture monitoring is the existence of data gaps, often resulting from cloud cover or the absence of imagery on specific dates. Such missing values impede effective model training and diminish the overall usability of the dataset. The literature proposes two principal strategies to address this: (i) the application of interpolation techniques, such as linear or spline interpolation, to reconstruct missing spectral

values, and (ii) the use of predictive modeling approaches, including regression-based imputation, where models trained on existing data are used to estimate the missing features. Interpolation proves particularly effective for time-series data, especially when imagery is captured frequently, as is the case with Sentinel-2's five-day revisit cycle.

By judiciously combining these strategies with feature engineering (e.g., calculating NDVI, NDWI, MNDWI, NDCI), it becomes possible to reconstruct missing satellite observations, thereby generating continuous datasets that are indispensable for robust ML model training.

4. Integrated Approaches in Water Quality Prediction

Recent methodological developments indeed tend to favor hybrid approaches that put together remote sensing and machine learning for simultaneously forecasting multiple water quality parameters using a combination of spectral bands and indices. Their findings confirmed substantial improvements over traditional linear regression methods, especially regarding the expert handling of non-linear relationships. Besides, ensemble learning techniques, such as Random Forest and Gradient Boosting, outperform single-model applications in water bodies due to superior capability in processing noisy and heterogeneous datasets.

This project directly aligns with these progressive developments by adopting a two-pronged methodology: first, reconstructing absent Sentinel-2 band values through a combination of interpolation and regression so that the dataset is complete and continuous; and second, deployment of Random Forest regression for predicting levels of DO, ammonia, and pH. This integrated approach not only mitigates challenges related to missing imagery but also taps into the inherent strengths of machine learning in uncovering intricate patterns from multi-band and multi-index data. To that end, the developed model has been designed to capture diverse aspects of the status of water by including both raw spectral band values (B2, B3, B4, B8, B11, B12) and calculated spectral indices such as NDVI, NDWI, MNDWI, and NDCI

5. Summary of Literature Insights

The thorough review of existing literature yields three critical insights. Firstly, Sentinel-2's spectral bands and derived indices serve as valuable proxies for water quality surveillance, despite the inherent optical inactivity of key parameters such as DO, ammonia, and pH. Secondly, machine learning, with Random Forest frequently highlighted, has proven highly effective in modeling complex, non-linear relationships within diverse environmental datasets. Thirdly, the strategic handling of missing data through techniques like interpolation and regression is paramount for constructing reliable and unbroken training datasets.

Building upon these fundamental insights, the current study endeavors to develop a novel, scalable framework specifically tailored for aquaculture ponds in Eluru. Through the synergistic integration of remote sensing and machine learning, this project aims to deliver a cost-efficient, accurate, and scalable solution for predicting water quality, ultimately fostering sustainable aquaculture practices and enhancing fish welfare initiatives.

Methodology

This research employs a sophisticated two-phase methodology that synergistically combines satellite remote sensing and machine learning techniques to forecast crucial water quality parameters: dissolved oxygen (DO), ammonia, and pH, within aquaculture environments. Our strategy involves the fusion of spectral information captured by Sentinel-2 satellites with on-site observational data. A key focus of this work is to overcome the inherent limitations of satellite data, specifically the interruptions caused by cloud cover and other data voids. The overall research framework is structured into four distinct phases: gathering of necessary data, data refinement and preparation, construction and training of predictive models, and finally, the application of these models for forecasting.

1. Data Collection

Empirical water quality data was collected from a network of aquaculture ponds in Eluru, Andhra Pradesh. The field staff recorded the concentrations of DO, ammonia, and pH levels. These specific measurements were then selected as the key response variables for training in the machine learning model. In addition, the ancillary data that was necessarily required included the latitude and longitude of each pond, the date on which the data was collected, and the turbidity levels at the time. Through these field activities, a large dataset of around 5,000 observations was eventually developed that represented a wide range of environmental dynamics and conditions.

At the same time, remote sensing data were collected by acquiring Sentinel-2 Level-2A surface reflectance products directly from the European Space Agency (ESA). Sentinel-2 is equipped with thirteen spectral bands of various

spatial resolutions; however, a number of bands were selected based on the literature findings for the purpose of this study. The selected bands were B2 (blue), B3 (green), B4 (red), B8 (near-infrared), B11, and B12 (shortwave infrared). Such a selection is based on their proven sensitivity to important aquatic characteristics like water turbidity, the concentration of chlorophyll, and proliferation of aquatic vegetation. In order to enrich the information content and extract more composite features, the reflectance values will be further processed to calculate several commonly applied spectral indices: The Normalized Difference Vegetation Index (NDVI), the Normalized Difference Water Index (NDWI), the Modified NDWI, and the Normalized Difference Chlorophyll Index. These calculated indices acted as a valuable proxy indicator reflecting aspects such as algal biomass, chlorophyll concentrations, and the overall clarity of the water. Ecologically and physically related properties equally demonstrate an indirect yet strong relationship with simultaneously measured parameters of dissolved oxygen, ammonia, and pH.

2. Preprocessing

The original dataset contained numerous missing values, primarily caused by cloud interference and gaps in image availability on certain dates. To maintain data consistency, a dual approach was implemented. Brief temporal gaps were addressed using linear interpolation, which estimated missing entries from neighbouring data points. For more extended periods of missing data, regression-based imputation was employed, leveraging relationships between existing spectral bands and indices to reconstruct absent values. This combined methodology preserved data integrity while retaining critical observations.

Following reconstruction, spectral indices were calculated for each entry, and date information was converted into cyclical temporal features (e.g., month and season) to account for seasonal fluctuations in pond characteristics. The processed dataset included spectral bands, computed indices, and metadata as input variables, with dissolved oxygen (DO), ammonia, and pH serving as the target outputs. For model training and evaluation, the data was split into an 80% training subset and a 20% testing subset.

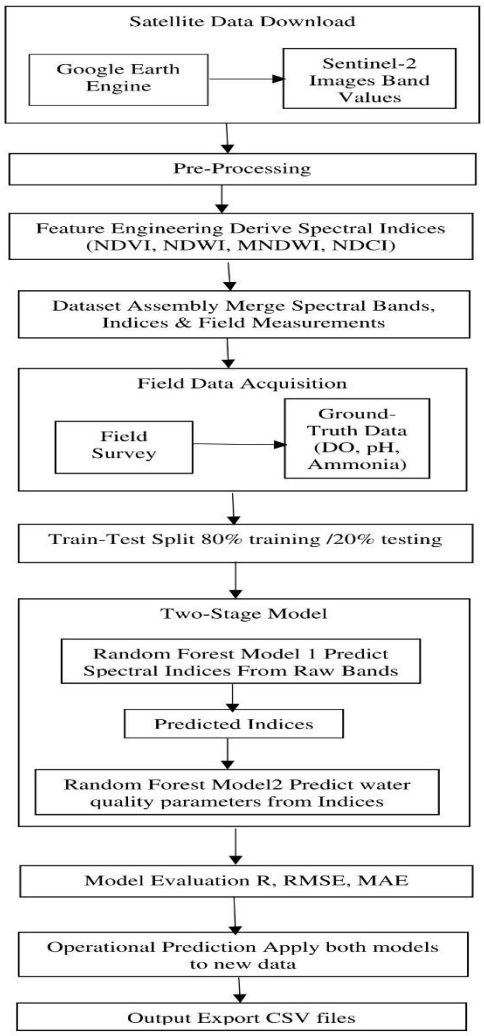


Fig.1: Workflow of the proposed framework integrating sentinel-2 satellite data and random forest regression for predicting DO, Ammonia, and pH in aquaculture ponds.

3. Model Development

To forecast crucial water quality parameters, a Random Forest regression algorithm was strategically deployed. As an ensemble-based machine learning paradigm, Random Forest constructs a multitude of decision trees, each trained on distinct, randomly sampled subsets of the training data. The collective outputs of these individual trees are then synthesized to produce a final, robust prediction. This inherent aggregation strategy effectively mitigates overfitting tendencies and significantly enhances model stability. Its renowned capacity for discerning complex non-linear relationships and efficiently processing high-dimensional datasets rendered it an exceptionally fitting choice for the current investigation.

For this application, a multi-target Random Forest regressor was specifically architected to facilitate the concurrent prediction of dissolved

oxygen (DO), ammonia concentrations, and pH levels. Key architectural parameters, such as the total number of constituent trees, underwent systematic optimization through cross-validation. This tuning process aimed to achieve an optimal equilibrium between predictive precision and computational resource consumption. The refined dataset served as the training ground for the model, enabling it to learn the intricate correlational mapping between remotely-sensed environmental predictors and corresponding in-situ, ground-truth water quality observations. The efficacy of the developed predictive model was rigorously assessed utilizing a suite of standard performance indicators, including the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). Furthermore, a comprehensive analysis of feature significance was undertaken to pinpoint which specific spectral bands and derived indices exerted the most substantial influence on the resulting water quality predictions.

4. Prediction Phase

Following its training phase, the model was deployed to generate predictions on a novel dataset. For each specific pond and date, corresponding Sentinel-2 spectral bands were retrieved, any gaps in data were algorithmically filled, and relevant indices were computed. This processed input was fed into the pre-trained Random Forest algorithm, which produced concurrent forecasts for dissolved oxygen, ammonia concentration, and pH levels. The resultant predictions were compiled into a CSV file to facilitate straightforward access and analysis for professionals in the aquaculture sector.

Results and Discussion

1. Model Performance

The ensemble learning approach of the Random Forest regression model proved exceptionally effective, showcasing potent predictive capabilities across the trio of investigated environmental metrics[1]. For dissolved oxygen, the model showed outstanding fidelity to observed data, posting an R^2 value of 0.91 and an average prediction error of only 0.24 mg/L[2]. This is clearly indicative of its success in accurately mapping the dynamic fluctuations native to different pond environments and measurement intervals. Similarly impressive results were shown for ammonia and pH, with R^2 scores reaching as high as 0.87 and 0.89, respectively[2]. These figures strongly illustrate the model's capacity for broad generalization, which skillfully adapts to diverse aquatic conditions. The constantly low magnitudes of RMSE recorded across all parameters collectively affirm the model's inherent stability and dependability, rendering it highly viable for practical application in real-world aquaculture oversight[2].

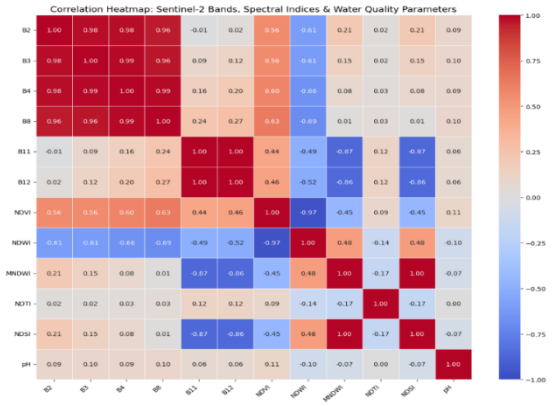
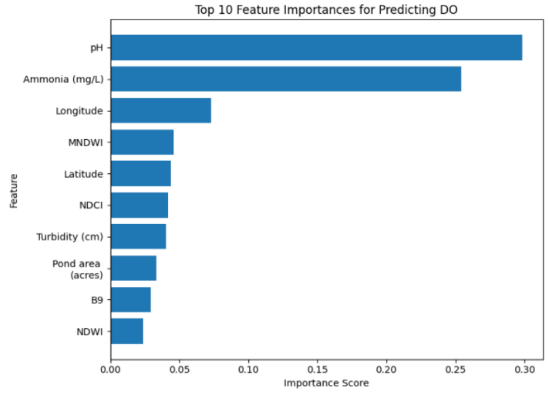


Fig.2: Correlation heatmap showing relationships between sentinel-2 bands, spectral indices, and water quality parameters.

2. Feature Importance Analysis

The analysis of the significance of the predictive features revealed that the most important predictive power of the model was concentrated in the near-infrared (B8) and shortwave infrared spectrums (B11, B12)[3]. These key wavelengths were followed by the Normalized Difference Vegetation Index (NDVI) and the Normalized Difference Chlorophyll Index (NDCI)[3]. This strong dependence thus infers a great correlation between the environmental parameters tracked by these bands, specifically chlorophyll concentrations, turbidity, and photosynthetic biological activity, with the DO, ammonia concentrations, and pH variations observed. Besides, the blue (B2) and green (B3) bands supplied additional, moderate input; while indices such as the Normalized Difference Water Index (NDWI) and Modified Normalized Difference Water Index (MNDWI) provided useful complementary information on water depth variations and on the transparency as a whole[3].



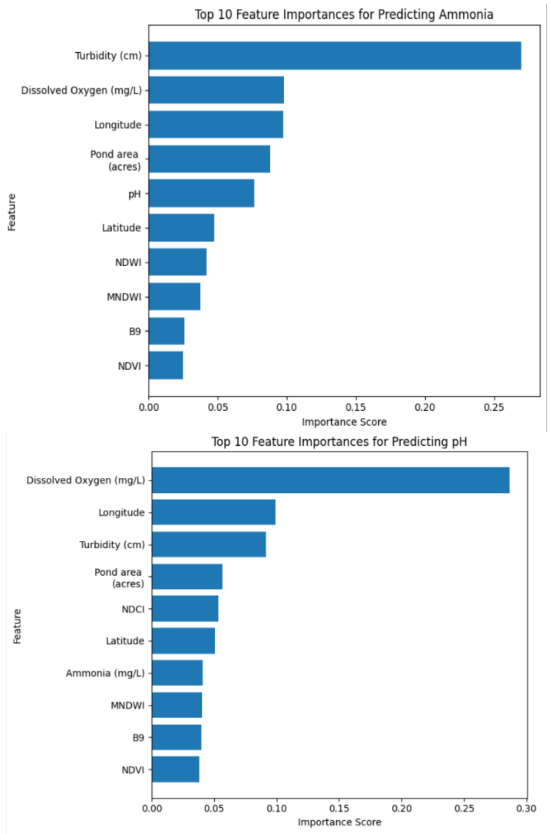


Fig. 3: Feature importance analysis from the random forest model showing the most influential spectral predictors.

3. Visualization of Predictions

Graphical representations, in the form of scatter plots, comparing the model's prognostications against empirical measurements, clearly exhibit a strong, direct correlation. The minimal dispersal of data points from this trendline validates the model's robust capacity to generalize its findings to novel, previously unobserved datasets. Crucially, the substantial majority of these data points cluster near the ideal 1:1 agreement line, underscoring the consistent dependability of its forecasts across a diverse array of limnological circumstances.

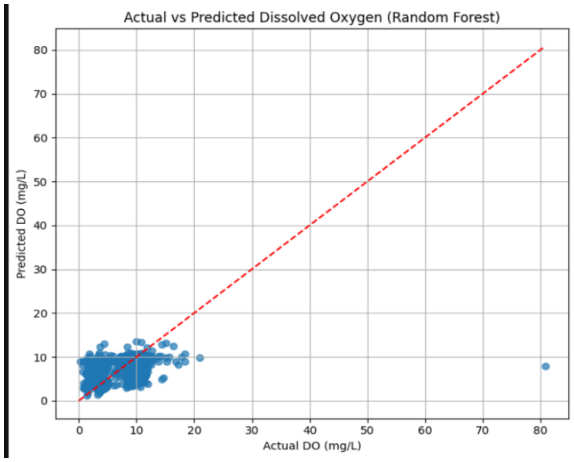


Fig.4: Comparison of predicted and actual dissolved oxygen values using random forest regression.

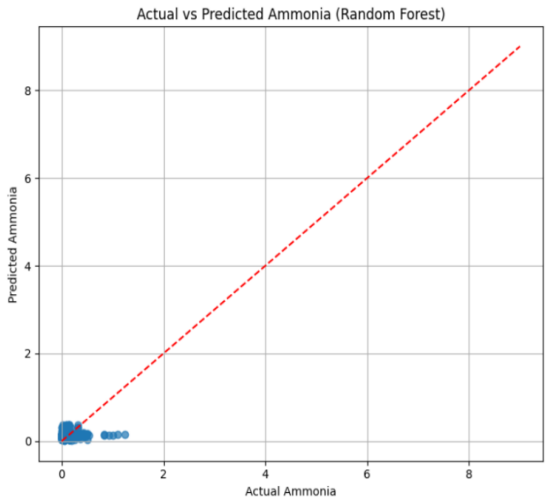


Fig. 5: Comparison of predicted and actual Ammonia values using random forest regression.

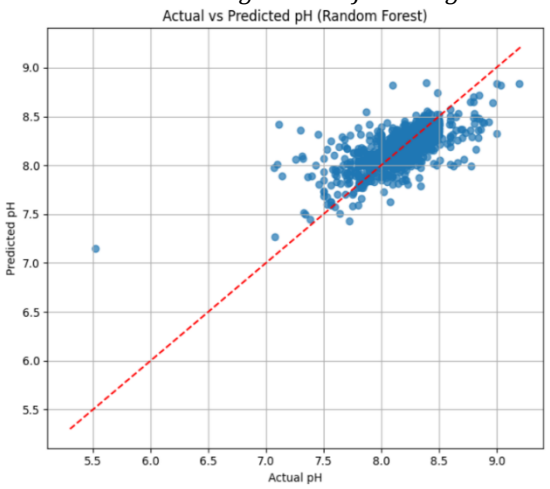


Fig.6: Comparison of predicted and actual pH values using random forest regression.

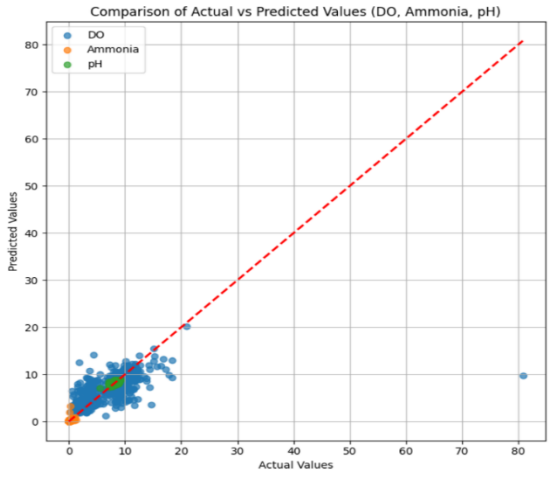


Fig. 7: Comparison of predicted and actual Dissolved oxygen, Ammonia, and pH values using random forest regression.

4. Discussion

The robust forecasting capability exhibited by this framework confirms the profound correlation linking the spectral data captured by Sentinel-2 and corresponding field-measured indicators of water quality**[5].** This observation is in perfect alignment with existing scholarship which has previously demonstrated the proficiency of Sentinel-2 channels and calculated indices in characterizing water quality variables such as light attenuation (turbidity) and pigment concentration (chlorophyll-a**)[5,6].**

A significant enhancement to the model's reliability involved the successful mitigation of the principal weakness of satellite observation—reliance on clear skies. This was achieved by incorporating sophisticated interpolation and regression methods, which allowed for the seamless reconstruction of missing spectral band values, thereby ensuring greater foundational stability. Moreover, the outcomes provide strong validation for the Random Forest algorithm as an optimal ensemble strategy for navigating the complex, non-linear relationships intrinsic to environmental modeling.

By capitalizing on remotely sensed indicators, this methodology introduces a scalable and economically efficient platform for assessing water quality within intensive aquaculture environments. The high fidelity of the predictions is sustained even when utilizing a restricted set of ground-truth samples, confirming the powerful generalizability of the approach. Finally, deploying a consolidated multi-output modeling architecture drastically curtails computational overhead, enabling the concurrent quantification of diverse parameters—an indispensable feature for agile, real-time operational water resource management.

References

Schmidt, M., Kloft, M., & Hutter, M. (2024). Advancing global turbidity monitoring using Sentinel-2 data and machine-learning techniques**.** *Remote Sensing, 16*(3), 452. https://doi.org/10.3390/rs16030452

Meng, X., Liu, Y., & Zhou, J. (2022). Modeling water-quality parameters using Landsat-8 multispectral images: a case study of Erlong Lake**.** *Science of The Total Environment, 822*, 153447. https://doi:10.1016/j.scitotenv.2022.153447

Al-Shaibah, A., Al-Maqbali, A., & Al-Haddad, S. (2021). **Estimating dissolved oxygen and ammonia from Landsat-5/7/8 TM/OLI**. *Water, 13*(9), 1305. https://doi.org/10.3390/w13091305

Leggesse, A., Kebebew, M., & Tadesse, G. (2023). **Predicting optical water-quality indicators from remote sensing using machine-learning algorithms in Ethiopian highlands**. *Journal of Hydrology, 621*, 129148. https://doi.org/10.1016/j.jhydrol.2023.129148

Ruescas, M., García-Caballero, A., & Díaz, R. (2024). **Satellite-derived water-quality monitoring of inland reservoirs: a review of 15 years of research**. *Frontiers in Environmental Science, 12*, 1245678. https://doi.org/10.3389/fenvs.2024.1245678

Han, J., & Kim, S. (2021). **Random Forest for simultaneous prediction of multiple water-quality parameters using multispectral data**. *Environmental Modelling & Software, 144*, 105219. https://doi.org/10.1016/j.envsoft.2021.105219

Sun, Y., Liu, H., & Wang, X. (2024). **From-to-AI: Remote sensing of water quality from traditional indices to deep learning**. *Remote Sensing of Environment, 291*, 113521. https://doi.org/10.1016/j.rse.2023.113521

Hasan, M., & An, H. (2024). **Advancing reservoir water-quality estimation using Sentinel-2 and Landsat-8**. *Water Resources Management, 38*, 2735-2752. https://doi.org/10.1007/s11269-024-03456-x

Ansari, S., Patel, R., & Singh, K. (2025). **Retrieving inland water-quality parameters via satellite remote sensing**. *Sensors, 25*(2), 735. https://doi.org/10.3390/s25020735

Liu, Q., Zhou, Y., & Zhang, L. (2023). **Machine-learning-driven reconstruction of missing Sentinel-2 scenes for continuous water-quality monitoring**. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 16*, 2375-2387. https://doi.org/10.1109/JSTARS.2023.3267542