Grenze International Journal of Engineering and Technology, June Issue

Water Quality Prediction of Puzhal Lake using Satellite Imagery and Machine Learning

M Geet Tej Mahesh[1], Kothuri Sai Venkata Dinesh[2], Sangapu Sreenivasa Chakravarthi[3] and S Sountharrajan[4]
[1-4]Department of CSE, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Chennai, India
Email: tejmahesh@outlook.com, saidinesh2567@gmail.com, ss_chakravarthi@ch.amrita.edu, s sountharrajan@ch.amrita.edu

***Abstract*— Water quality is one of the most important components of environmental sustainability and human health, but conventional monitoring is often insufficient to provide the spatial and temporal resolution needed for effective management. This paper proposes a new approach to predicting water quality in Puzhal Lake based on the integration of high-resolution Sentinel-2 satellite images and in-situ data provided by the Tamil Nadu Water Board from 2019 to 2023. The development of the proposed methodology entailed a series of steps that included pre-processing of the satellite data to remove noise, mask clouds, correct atmospheric effects and perform spatial-temporal integration to generate high definition and accurate images. Some spectral ratios like the blue to green ratio and the Normalized Difference Water Index (NDWI) were also extracted to see the subtle changes in water quality. These indices are combined with the ground data of DO, pH, Turbidity, and TDS to form a single dataset. Several machine learning algorithms are developed to predict the water quality parameters namely Random Forest, Gradient Boosting and Artificial Neural Networks. The models are trained and optimized to the best of their abilities through hyperparameter tuning and cross validation and give a good performance with RMSE as low as 0.078 and R^2 of 0.911. Besides contributing to the improvement of current practices in water resource management, this study also offers practical recommendations for policy making and environmental protection in the context of sustainable development.**

***Index Terms*—Water quality; Environmental sustainability; Sentinel-2; In-situ data; Puzhal Lake; Tamil Nadu Water Board; Spectral ratios; Normalized Difference Water Index (NDWI); Machine learning; Hyperparameter tuning.**

I. INTRODUCTION

Water is becoming an increasingly essential resource of prime importance to all environmental aspects, human health, and the economy itself. Historical urbanization, industrialization, and agricultural activities have invaded the waters of several natural water bodies. Conventional monitoring, which relies on sporadic in-situ sampling and laboratory analysis, sometimes does not reflect the highly complex spatial and temporal variations in the behaviour of large lakes. Therefore, the goal of this study is to propose a novel framework through which water quality prediction for Puzhal Lake can be made by means of advanced remote sensing techniques, extensive ground-truth measurements, and advanced machine learning algorithms. A pre-processing pipeline accomplishes thorough noise reduction, cloud masking, atmospheric correction, and spatial-temporal data fusion based on the input high-resolution Sentinel-2 satellite imagery. The resulting data, on its own and through processing, provide reliable and consistent data for long periods.

Grenze ID: 01.GIJET.11.2.637

The Tamil Nadu (India) Water Board has collected in situ measurements with time limits from 2019 to 2023 on important parameters like Dissolved Oxygen (DO), pH, Turbidity, and Total Dissolved Solids (TDS) connected to satellite-derived data. The extraction of meaningful spectral indices, including the blue-to-green ratio and Normalized Difference Water Index (NDWI), helps the method capture subtle variations detailed as indicators of water quality. So these remote sensing indices with accurate in situ measurements produce an excellent multidimensional data source to empower our predictive modelling.

The central challenge the work strives to address is to derive a tool for real-time monitoring with great scalability, ensuring it can address not just the intrinsic limitations of contemporary water quality monitoring but also offer usable insight into management of water resources. These challenges become compounded as atmospheric disturbances, sensor noise, and temporal offsets between the overpassing of the satellite and in situ sampling need to be reduced. To address these challenges, we have designed a complete data integration approach for aligning satellite imagery with ground measurements using a precisely defined temporal tolerance so that the resultant dataset will truly capture the water quality condition at any time instant. Our approach also capitalizes on the latest machine-learning techniques to model the highly non-linear relationships governing the dynamics of water quality parameters. Such methodologies shall include training the randomly generated models-Random Forest, Gradient Boosting, and ANN-based models-systematically evaluate and optimize their predictive performance using metrics like Root Mean Square Error (RMSE) and the coefficient of variation.

This study is dedicated to developing a scalable, real-time water-monitoring system that complements and overcomes the limitations of traditional water quality monitoring and provides user-relevant actionable information for efficient resource management. Such application becomes complicated by the need to record measurements with minimal interference from atmospheric disturbances, sensor noise, or temporal mismatches that exist with many even the best water bodies during the time intervals between their overpassing with satellites and ground sampling. In order to address these requirements, we have established a comprehensive data-joining strategy where satellite imagery is fused with ground measurements using carefully defined temporal tolerance boundaries, resulting in a dataset that is truly representative of the actual water quality condition at any moment in time. Furthermore, leveraging advanced machine learning methods, our approach allows for modelling the complex non-linear dependency structures that govern the water quality dynamics. Furthermore, for evaluating the accuracy of prediction models, we will make comparative analyses among different models - Random Forest, Gradient Boosting, and Artificial Neural Networks (ANN) systematically training them and optimizing their performance using metrics such as Root Mean Squared Error (RMSE), coefficients of determination, and residuals analysis.

II. RELATED WORKS

Water quality monitoring has therefore gained increased research importance due to its dire consequences for sustainable environmental and public health. The rapid development of remote sensing technologies with the development of machine learning tools opened up new avenues for water quality assessment. Landsat and Sentinel imagery, spectral analysis, and predictive modelling were utilized to monitor and estimate water quality parameters, with all three approaches providing significant understanding that have informed our present work.

Early efforts in this subject were focused on remote sensing of turbidity utilizing high-resolution satellite imagery. For example, Ma et al. [1] researched the application of Sentinel-2 image data to estimate turbidity in Northeast China lakes using machine learning algorithms to relate the spectral indices with in-situ measurements. Their approach suggested the spectral indices, notably the ones sensitive to water constituents, might serve as potent proxies for water quality monitoring. Certainly, this work greatly formed the basis of subsequent investigations, showcasing that Sentinel-2 images can delineate salient water quality indicators across extensive spatial extents.

Building on the above, Adusei et al. [2] applied remote sensing technologies to predict and map water quality within the Owabi Reservoir. This methodology fused satellite imageries with machine learning models for the generation of detailed water quality maps communicating reservoir spatial variability. This study further highlighted the importance of combining remotely sensed data with ground-truth observations to maximize water quality prediction accuracy. Zheng et al. [3] have similarly illustrated the integration of hydrodynamic modelling with deep learning approaches in the prediction of harmful algal blooms in large water bodies. With an innovative framework allowing for the dynamic forecast of algal bloom events, this particular concern bears direct association with the overall water quality.

In addition, Fuentes-Aguilera et al. [4] researched the recovery of time series data of water volume in Lake Ranco (South Chile) from satellite altimetry. They found that variations in water volume were related to climatic phenomena, showing that satellite data could successfully be used for monitoring hydrological changes through time. This temporal analysis is fundamental to our four-year study and one that aims to capture seasonal water quality change. Similarly, Wang et al. [5] devised an urban water extraction methodology that integrated deep learning technologies with Google Earth Engine, providing a scalable solution for water resource management in urban situations. Their strategy not only demonstrated the applicability of deep learning in processing large volumes of satellite data but also emphasized the need for automated workflows in environmental monitoring.

In addition, spectral analyses have been a prominent research concentration area, with examples such as Rodr´ıguezLopez et al. [6] whose work employed LANDSAT images´ to monitor chlorophyll-a concentration in the Lake Laja in Chile. It further elucidated a better understanding of algal concentrations, an indicator for water quality based on spectral characteristics. Rodr´ıguez Lopez et al. [7] used advanced´ machine-learning algorithms to examine water quality in Lake Llanquihue, Southern Chile through estimation of its parameters. They performed a comparative analysis on different types of modeling methodologies, hence corroborating the ground that advanced machine learning algorithms can complement the accuracy of prediction related to the assessment of water quality.

The challenge of incompleteness or lack of data in remote sensing has been approached by several researchers. Thus, Siabi et al. [8] proposed a spatio-temporal gap filling algorithm which was used to reconstruct data from the remotely sensed precipitation, land surface temperature (LST), and evapotranspiration (ET) datasets. It adds to the continuity and reliability of the dataset which is important when integrating satellite data with in-situ observations. Removal and detection of clouds is also part of pre-processing processes of satellite imagery. Shao et al. [9] developed cloud detection algorithm with multiscale feature using convolutional neural networks (CNNs). On the other hand, Ma et al. [10] applied deep learning model using cloud-matting technique for cloud removal. Such techniques will be very handy in ensuring the satellite imagery to be mostly artifact free during the assessment of water quality, thus, improving the accuracy of the achieved spectral indices.

While the insights from these studies are invaluable in terms of methodologies and results, our study is significantly different in that it focuses on Puzhal Lake, which has not been much represented in the literature integrating remote sensing with machine learning. From 2019 to 2023, we analyse water quality across both spatial and temporal domains using a very robust dataset. In view of the scheme that integrates satellite remote sensing with in-situ measures, these challenges include atmospheric effects, temporal mismatches, and the very nonlinear nature of water quality dynamics. In addition, the applications presented consider with high detail hyperparameter tuning, cross-validation, and spatial-temporal visualizations to make certain that our predictive scheme works accurately and on an allowable scale.A synthesis of the studies thus far identified several research gaps which our work seeks to fill.

In summary, the literature indicates that while there have been considerable improvements in remote sensing and machine learning for water quality monitoring, there is still a requirement for integrated multi-year studies combining high resolution satellite imagery with detailed field appraisal of the water quality. Our research extends these existing ones by presenting a comprehensive framework especially designed for Puzhal Lake. This integrated approach improves the accuracy of predicting water quality and supplies a scalable and real-time monitoring tool, vital for successful environmental management. Therefore, our work builds on past findings and offers a rare voice and a major leap forward for water quality monitoring.

III. METHODOLOGY

A. Data Acquisition

The foundation is set on two fundamental sets of data collection. First is the 2019–2023 high-resolution Sentinel2 satellite data over Puzhal Lake (coordinates roughly 80.1283°E–80.1969°E, 13.1340°N–13.2029°N). Downloaded images on Google Earth Engine include multispectral information required for water quality parameter extraction. The second piece of information is in-situ monitoring of water quality, i.e., Dissolved Oxygen (DO), pH, Turbidity, and Total Dissolved Solids (TDS), by the Tamil Nadu Water Quality
Board on a monthly basis.

B. Preprocessing Pipeline

Satellite imagery passes through an aggressive pre-processing stream for enhancing data quality and homogeneity prior to being injected into the system. Occlusions and artifacts are eliminated through a sequence of noise reduction operations and cloud mask procedures. Cloud mask is a threshold-based process wherein pixels of low reflectance in the blue and green bands (e.g., ¡0.1) were identified and eliminated. The imagery is subsequently atmospherically corrected with standard correction models such that reflectance values become time-comparable. Spatial cropping also downsizes each of the images to a uniform resolution of 256×256 pixels, the extent of Puzhal Lake's size. In case of low brightness (mean reflectance below some specified threshold), histogram equalization is used for enhanced contrast. Additional acceptable variation is introduced into the data by using data augmentation techniques like rotation (+20° or -20°), zoom, and spatial shifting.

Fig. 1. Satellite (Sentinel-2) Overview of Puzhal Lake

- Feature Engineering and Data Integration: Following pre-processing, key spectral indices are derived from the satellite imagery. Notably, the blue-to-green ratio is computed using:

$$R_{bg} = \frac{B_2}{B_3 + \varepsilon} \tag{1}$$

where B_2 and B_3 represent the reflectance in the blue and green bands, respectively, and ε is a small constant (e.g., 10^{-10}) to avoid division by zero. Additionally, the Normalized Difference Water Index (NDWI) is calculated as:

$$\text{NDWI} = \frac{\text{Green} - \text{SWIR}}{\text{Green} + \text{SWIR} + \varepsilon} \tag{2}$$

At the same time, the satellite data matches in time with in-situ measurements through a three-day tolerance window to obtain the most representative ground measurement for each satellite observation.

C. Machine Learning Model Development

This integrated dataset provides the basis for predictive modelling. This means the dataset is divided directly into testing and training sets in an attempt to estimate the model's generalizability.
Three machine learning models to be considered are as follows:

- Random Forest (RF): An ensemble of decision trees to model non-linear relationships.
- Gradient Boosting (GB): Prediction is improved sequentially with emphasis on the previous errors.
- Artificial Neural Network (ANN): MLPRegressor whose network architecture involves two hidden layers (each carrying 50 neurons) and ReLU activation functions.

During training, the models are taught to tune parameters in a manner that minimizes predictions' Mean Squared Error (MSE) against actual measures. Test results were presented in terms of RMSE values and coefficient of determination (R^2) for the models.

D. Hyperparameter tuning and evaluation

The predictive accuracy of each model is therefore enhanced by hyperparameter tuning that employs techniques like GridSearchCV. Important parameters for Random Forest include the number of trees and the maximum depth; Gradient Boosting and ANN models are adjusted mainly on the number of estimators and learning rates, with ANN model architecture also being a focus for tuning. Cross-validation ensures the given parameters generalize well to the different data split and mitigate overfitting risk during this stage. The hyperparameter tuning stage itself gives rise to a choice of configurations that have excellent achievement, with our best model (ANN) providing an RMSE of 0.078 and an R-squared value of 0.911.

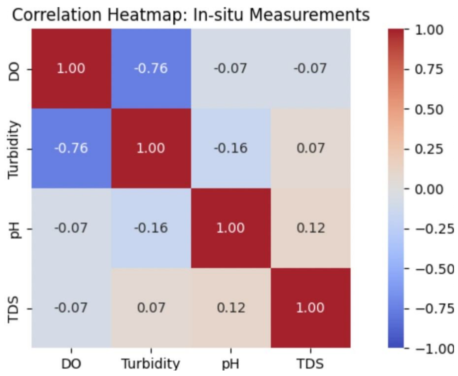

Fig. 2. Correlation Heatmap of parameters

E. Spatial-Temporal Analysis and Visualization

The last step in our methodology entails detailed spatial and temporal analyses to interpret model outputs and appraise the dynamics of water quality. Time series analyses are performed on the unified dataset by resampling over time (such as taking monthly averages) to highlight seasonal trends and anomalies of major water quality parameters. Spectral indices (for example, NDWI and TDS) are overlaid on a map of Puzhal Lake to visualize spatial heterogeneity and distinguish areas with markedly different water quality characteristics. These insights are further mapped using correlation heatmaps, pair plots, and boxplots to establish relationships among the different parameters. Such visual analytics would further support model validation and provide management recommendations regarding the water resources.

IV. RESULTS

A. Model Performance Evaluation

The predictive models were developed using the test dataset. The performance results show that RMSE value with ANN is 0.078 and R^2 is 0.911, which is better than Random Forest and Gradient Boosting RMSE of 0.082 and 0.085 and R^2 of 0.900 and 0.894, respectively. This indicates that the ANN model is good at modelling the non-linear relationship in the fused data given in this work.

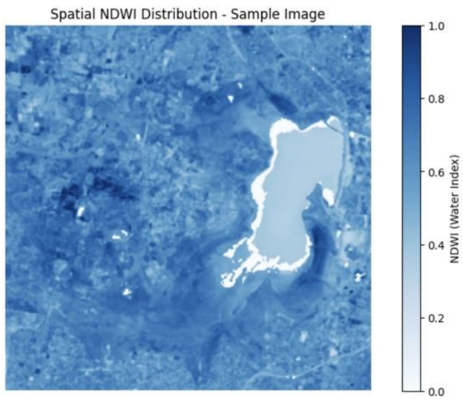

Fig. 3. Spatial NDWI View of Puzhal Lake

TABLE I: MODEL PERFORMANCE

Model	RMSE	R_2
Random Forest	0.082	0.900
Gradient Boost	0.085	0.894
ANN	0.078	0.911

B. Spatial Analysis Result

Spatial analysis was carried out by mapping spectral indices-such as NDWI and TDS estimates-into the geographic area of Puzhal Lake. These maps generated indicated large spatial heterogeneity, where some areas

have consistently low water quality. The visual patterns coincide with known pollution sources and ambient variability across the lake. heterogeneity, where some areas have consistently low water quality. The visual patterns coincide with known pollution sources and ambient variability across the lake.
Time series plots of water quality parameters (DO, pH, Turbidity, and TDS) were derived by resampling the unified dataset to monthly averages. The plots, therefore, clearly represent seasonal patterns by which DO levels vary with seasonal climatic changes, and turbidity shows periodic increments during rainfall events. These temporal patterns are corroborated with both in-situ observations and historic environmental records.

C. Temporal Analysis Results

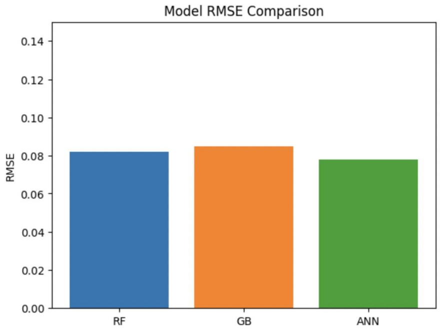

Fig. 4. RMSE Comparison of Models

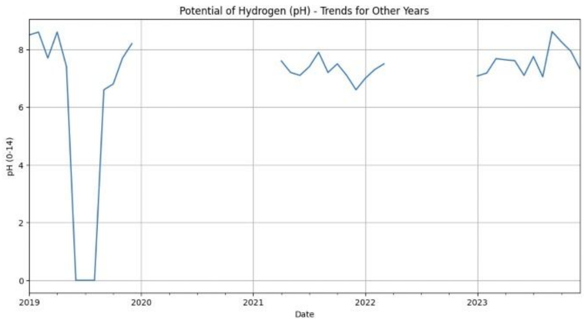

Fig. 5. Predicted pH Trends Over the Years

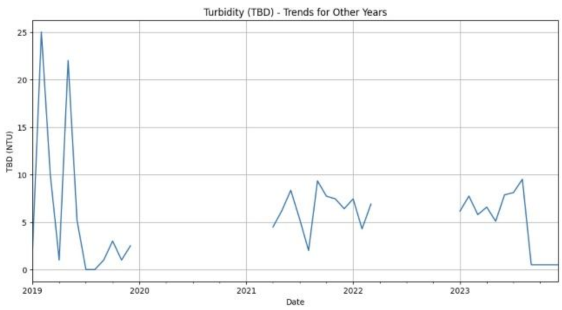

Fig. 6. Predicted Turbidity Trends Over the Years

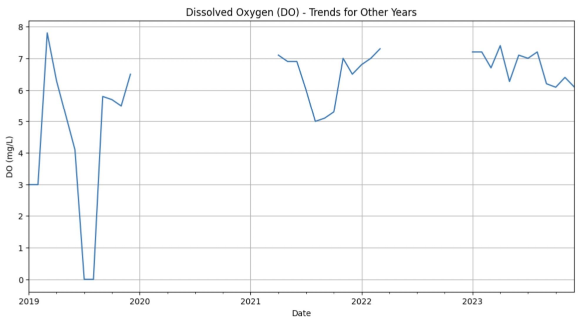

Fig. 7. Predicted Dissolved Oxygen Trends Over the Years

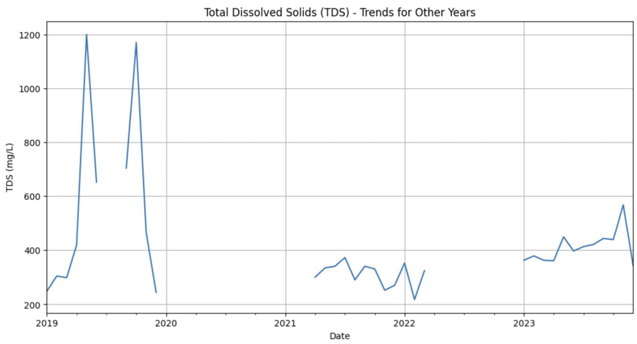

Fig. 8. Predicted TDS Trends Over the Years

D. Statistical Analysis and Validation

Further statistical analyses were carried out to ascertain the robustness of the predictive models, including calculating the mean absolute errors and plotting the distribution of errors. The low variation of prediction errors across different cross validation folds indicates a stable and robust model.

E. Discussion

Experimental results confirm that the given integrated framework performs better than current state-of-the-art remote sensing water quality estimation with in-situ data fusion and machine learning. Our approach provides the following over traditional monitoring techniques:

- ➢ Increased Predictive Precision: Our models precisely forecast complex, non-linear relationships between parameters of water quality with a minimum RMSE of 0.078 and R^2 of as much as 0.911, forecasting crucial indicators such as DO, pH, Turbidity, and TDS correctly.
- ➢ Strong Spatial-Temporal Analysis: The correct spatial location and time series analysis exhibit obvious seasonality and spatial heterogeneity of Puzhal Lake and provide valuable evidence of localized water quality variation, and enable effective water resource management.
- ➢ Data Fusion and Pre-processing: The rigorous pre-processing workflow of noise elimination, cloud masking, atmospheric correction, and spatial-temporal fusion yields high-fidelity satellite imagery that, in the event of the integration of in-situ measurement, forms a reliable multidimensional dataset.
- ➢ Scalable Real-Time Monitoring: The design of the framework makes it scalable and vulnerable to real-time data fusion, and therefore ideal for continuous long term water quality monitoring and timely environmental management decision-making.

The above observations present proof of the ability of our method to continuously function on the basis of a large number of temporal sets of data and are therefore an important addition to conventional water quality monitoring.

V. CONCLUSION

The research adopts a novel, coupled methodology to water quality prediction in Puzhal Lake in which high-resolution Sentinel-2 satellite imagery is synergistically combined with in situ water quality observations gathered over four years from 2019-2023. A strict pre-processing workflow converts remote sensing data into precise and high-quality imagery by eliminating noise, masking clouds, correcting atmospherics, and spatiotemporal integration. Critical spectral indices such as blue-to-green ratio, Normalized Difference Water Index (NDWI) are calculated and combined with ground samples of Dissolved Oxygen, pH, Turbidity, and Total Dissolved Solids to form a strong multidimensional dataset that captures the spatial and temporal dynamics of water quality. Subsequently, the cutting-edge machine learning algorithms in the form of Random Forest, Gradient Boosting, and Artificial Neural Networks were trained and hyperparameter optimized along with cross-validation. The excellent performance of the ANN model, with low RMSE of 0.078 and high R^2 of 0.911, reflects the robustness of our method in being capable of capturing the intricate nonlinear relationships underlying the environmental data effectively. Temporal and spatial analysis, as well as visual outputs, identified clear seasonal and spatial patterns, thus enabling actionable decision-making in water resource management.
The observation hence intensified the intensity of water quality monitoring by providing an extendable real-time warning system oriented toward Puzhal Lake. The cooperative process makes the restraint possible with conventional monitors and creates opportunities for extended study in totally computerized multi-partial environmental monitoring systems.
Subsequent upgrades would seek to improve the data fusion process further by using more sophisticated deep learning architectures and increasing the framework’s usability on other bodies of water. The study opens the door to more enlightened decision-making in the field of environmental management, in accordance with meeting the current challenges and the future requirements in water quality evaluation.

ACKNOWLEDGMENT

The authors would also like to extend their sincere thanks to the helpful in-situ water quality data measurements taken by the Tamil Nadu Water Quality Board, which have made this study possible. Sincere appreciation to the respective faculty mentors and all colleagues of Amrita Vishwa Vidyapeetham, Chennai, for regular guidance and fruitful comments throughout this project. The authors also thank the GEE remote sensing team for assistance in obtaining and processing Sentinel-2 images. The efforts of all the above contributors have greatly enhanced the quality and worth of our research on Puzhal Lake water quality forecasting.

REFERENCES

[1] Ma, Y., Song, K., Wen, Z., Liu, G., Shang, Y., Lyu, L., ... & Hou, J. (2021). Remote sensing of turbidity for lakes in northeast China usingSentinel-2 images with machine learning algorithms. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 9132-9146.

[2] Adusei, Y. Y., Quaye-Ballard, J., Adjaottor, A. A., & Mensah, A. A. (2021). Spatial prediction and mapping of water quality of Owabi reservoir from satellite imageries and machine learning models. The Egyptian Journal of Remote Sensing and Space Science, 24(3), 825833.

[3] Zheng, L., Wang, H., Liu, C., Zhang, S., Ding, A., Xie, E., ... & Wang, S. (2021). Prediction of harmful algal blooms in large water bodies using the combined EFDC and LSTM models. Journal of Environmental Management, 295, 113060.

[4] Fuentes-Aguilera, P., Rodr´ıguez-Lopez, L., Bourrel, L., & Frappart, F.´ (2024). Recovery of Time Series of Water Volume in Lake Ranco (South Chile) through Satellite Altimetry and Its Relationship with Climatic Phenomena. Water, 16(14), 1997.

[5] Wang, Y., Li, Z., Zeng, C., Xia, G. S., & Shen, H. (2020). An urban water extraction method combining deep learning and Google Earth engine. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13, 769-782.

[6] Rodrıguez-Lopez, L., Duran-Llacer, I., Gonz´ alez-Rodrıguez, L., del Rio,´ R. A., Cardenas, R., Parra, O., ... & Urrutia, R. (2020). Spectral analysis´ using LANDSAT images to monitor the chlorophyll-a concentration in Lake Laja in Chile. Ecological Informatics, 60.

[7] Rodr´ıguez-Lopez, L., Bustos Usta, D., Bravo Alvarez, L., Duran-Llacer,´ I., Lami, A., Mart´ınez-Retureta, R., & Urrutia, R. (2023). Machine learning algorithms for the estimation of water quality parameters in Lake Llanquihue in Southern Chile. Water, 15(11), 1994.

[8] Siabi, N., Sanaeinejad, S. H., & Ghahraman, B. (2020). Comprehensive evaluation of a spatio-temporal gap filling algorithm: Using remotely sensed precipitation, LST and ET data. Journal of Environmental Management, 261, 110228.

[9] Shao, Z., Pan, Y., Diao, C., & Cai, J. (2019). Cloud detection in remote sensing images based on multiscale features-convolutional neural network. IEEE Transactions on Geoscience and Remote Sensing, 57(6), 4062-4076.

[10] Ma, D., Wu, R., Xiao, D., & Sui, B. (2023). Cloud removal from satellite images using a deep learning model with the cloud-matting method. Remote Sensing, 15(4), 904.

[11] Hu, K., Wang, Y., Feng, B., Wu, D., Tong, Y., & Zhang, X. (2020). Calculation of water environmental capacity of large shallow lakes–a case study of Taihu Lake. Water Policy, 22(2), 223-236.

[12] Uddin, M. G., Nash, S., Rahman, A., & Olbert, A. I. (2022, December). Development of a water qualit y index model-a comparative analysis of various weighting methods. In Mediterranean Geosciences Union Annual Meeting (MedGU-21). Istanbul (pp. 1-6).

[13] Gomez, D., Salvador, P., Sanz, J., & Casanova, J. L. (2021). A new´ approach to monitor water quality in the Menor sea (Spain) using satellite data and machine learning methods. Environmental pollution, 286, 117489.