Application of Machine-Learning Techniques for Water Quality Assessment in Coastal Environments: A Case Study of the Jacarepaguá Lagoon System at Rio de Janeiro/BR

Dannylo Cardoso Mauricio[1], Jader Lugon Jr[2], André Merlo[1], Mayara Omai[1], Pedro Henrique González[3], Raphael Guerra[4], Wagner Telles[4], and Diego Brandão[1(✉)]

[1] Federal Center for Technological Education of Rio de Janeiro, Rio de Janeiro, Brazil
{dannylo.mauricio,andre.merlo}@aluno.cefet-rj.br,
{mayara.omai,diego.brandao}@cefet-rj.br
[2] Federal Fluminense Institute, Niterói, Brazil
jaderlugon@iff.edu.br
[3] System Engineering and Computer Science Program, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
pegonzalez@cos.ufrj.br
[4] Federal Fluminense University, Niterói, Brazil
{rguerra,wrambaldi}@id.uff.br

Abstract. Coastal environments, such as beaches, canals, and estuaries, are ecologically vital, supporting diverse species and regulating the interface between continental and oceanic systems. Despite their importance, these ecosystems face growing threats from anthropogenic activities that compromise their environmental balance. In this context, data-driven approaches such as machine learning offer promising tools for monitoring and managing water quality. However, many existing studies focus on limited indicators and often overlook temporal and spatial dynamics. This study addresses these gaps by applying machine-learning models to assess water quality in the Jacarepaguá Lagoon System, a complex of four interconnected lagoons in Rio de Janeiro, Brazil. While the analysis centered on biochemical oxygen demand (BOD), future work will incorporate additional indicators such as turbidity and dissolved oxygen, and adopt temporal modeling strategies. The models developed achieved a mean squared error of 17$(mg/L)^2$, constrained by the sparsity of monitoring data. These findings underscore both the potential and limitations of current predictive approaches and highlight the need for more comprehensive and continuous data collection to support evidence-based environmental management.

Keywords: Estuaries · Water Quality · Machine learning · Computational Modeling

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
O. Gervasi et al. (Eds.): ICCSA 2025, LNCS 15650, pp. 295–306, 2025.
https://doi.org/10.1007/978-3-031-96962-1_20

1 Introduction

Coastal environments—such as beaches, estuaries, and lagoon systems—play a vital ecological and socio-environmental role by supporting diverse aquatic and terrestrial species and regulating hydrological dynamics between continental water bodies and the oceans [1–3]. These ecosystems are also crucial for human activities, including fisheries, tourism, and urban development. However, increasing anthropogenic pressures—such as industrial effluent discharge, untreated domestic sewage, and unplanned urban expansion—have resulted in severe environmental degradation [4,5]. Coastal water contamination reduces biodiversity, degrades water quality, and poses serious public health risks, undermining local communities and ecosystem resilience.

One of the most critical consequences of this degradation is eutrophication, driven by excessive nutrient inputs—particularly nitrogen and phosphorus—which can lead to harmful algal blooms (HABs), oxygen depletion, and the collapse of aquatic ecosystems [6]. This issue is especially prevalent in coastal lagoon systems, where rapid urbanization has intensified pollution in estuarine zones [7]. In this context, water quality monitoring is essential to assess pollution levels and guide mitigation strategies in alignment with the United Nations Sustainable Development Goals (SDGs)—notably Goal 6 (Clean Water and Sanitation), which promotes improved water quality and wastewater management, and Goal 14 (Life Below Water), which emphasizes the conservation and sustainable use of marine and coastal ecosystems [8].

Although accurate, conventional water quality assessment methods rely on periodic field sampling and laboratory analysis, which are resource-intensive, time-consuming, and often unable to provide real-time data [9]. Computational modeling has emerged as a complementary approach, enhancing environmental data analysis and enabling trend prediction [7,10–14]. In particular, machine learning (ML) techniques have demonstrated considerable potential in environmental applications by supporting data-driven decision-making and delivering more accurate forecasts [11,15].

This study evaluates and compares the performance of machine learning models for predicting water quality parameters in the Jacarepaguá Lagoon System (JLS)—a complex of interconnected lagoons located in Rio de Janeiro, Brazil. The methodology begins with an exploratory data analysis (EDA) to examine the distribution, trends, and interrelationships among key water quality variables. Subsequently, multiple ML models—including Artificial Neural Networks (ANN), Support Vector Regression (SVR), Decision Tree Regressor (DTR), and Random Forest Regressor (RFR)—are trained and evaluated using the Mean Squared Error (MSE) metric to quantify predictive accuracy. Results indicate that data scarcity substantially limits model performance, increasing error rates and reducing reliability. These findings highlight the critical importance of high-resolution, continuous environmental monitoring to support the development of accurate, ML-based predictive models for water quality management.

The remainder of this paper is structured as follows: Sect. 2 presents the theoretical framework, discussing water quality assessment methods and machine

learning applications. Section 3 details the methodology, including data collection, preprocessing, and model selection. Section 4 reports the results of the ML model comparison and discusses the findings. Finally, Sect. 5 concludes with final remarks and recommendations for future research.

2 Theoretical Framework

Assessing water quality in coastal environments is a multidisciplinary challenge that combines conventional environmental monitoring with computational intelligence to enhance predictive capabilities and decision-making. This section reviews the foundational concepts of water quality assessment, focusing on applying machine learning (ML) techniques in recent studies.

2.1 Water Quality and Environmental Impact

Coastal lagoons and estuarine systems are key in maintaining biodiversity and regulating hydrological cycles. However, these environments are particularly susceptible to anthropogenic pressures, such as untreated domestic and industrial effluents, unregulated urbanization, and land-use changes [5]. One of the most severe consequences of water pollution in such systems is *eutrophication*—a process triggered by excessive nutrient loading (notably nitrogen and phosphorus), leading to algal blooms, oxygen depletion, and the degradation of aquatic ecosystems [6].

Water quality is typically assessed through parameters such as Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), pH, turbidity, total phosphorus, and thermotolerant coliforms. These indicators provide critical insights into the level and sources of pollution and potential ecological risks [9].

The deterioration of coastal water bodies is a global concern addressed in the United Nations Sustainable Development Goals (SDGs). Specifically, **SDG 6 (Clean Water and Sanitation)** aims to improve water quality and expand access to wastewater treatment. In contrast, **SDG 14 (Life Below Water)** promotes the conservation and sustainable use of marine and coastal ecosystems [8].

2.2 Traditional Methods for Water Quality Monitoring

Traditional water quality monitoring relies on field sampling and laboratory analysis, often guided by environmental standards set by institutions such as the Instituto Estadual do Ambiente (INEA) and the National Environmental Council (CONAMA) [17–19]. Although these methods produce high-precision results, they are resource-intensive, limited in temporal resolution, and often unable to capture rapid changes in water conditions.

To address these limitations, computational models—particularly hydrodynamic models—have been employed to simulate physical, chemical, and biological processes in aquatic systems [20]. However, these models typically require large, high-quality datasets and substantial computational resources, limiting their real-time or large-scale monitoring applicability.

2.3 Machine Learning Techniques for Water Quality Prediction

Recent advances in Machine Learning have created new possibilities for improving water quality assessments' accuracy, scalability, and responsiveness. ML models can handle high-dimensional data, identify complex patterns, and provide robust predictions based on historical and real-time inputs [7]. Several ML techniques have been successfully applied to water quality modeling, including:

- **Artificial Neural Networks (ANNs)**: Inspired by biological neural systems, ANNs consist of interconnected neurons arranged in layers. These models are particularly effective in capturing non-linear relationships among variables and have been used for predicting BOD, DO, pH, and nutrient concentrations [9,16].
- **Support Vector Machines (SVMs)**: SVMs perform classification or regression tasks by finding optimal hyperplanes in feature space. Variants using kernel functions—such as linear, polynomial, and radial basis function (RBF)—can model non-linear interactions between environmental variables with high accuracy [7].
- **Decision Trees (DTs) and Random Forests (RFs)**: DTs are rule-based models that segment data into decision paths, making them intuitive and interpretable. RFs aggregate multiple DTs using bootstrap aggregation (bagging), improving robustness and reducing overfitting. These models are widely used in ecological modeling for predicting parameters such as turbidity and nutrient concentrations [7].
- **Gradient Boosting Machines (GBMs)**: GBMs are ensemble methods that build a sequence of weak learners (typically decision trees), where each new model attempts to correct the errors of the previous ones. Algorithms like XGBoost and LightGBM have shown excellent predictive accuracy in environmental modeling, especially in datasets with missing values and complex feature interactions.
- **k-Nearest Neighbors (k-NN)**: A non-parametric method based on proximity in feature space, k-NN has been applied in water quality classification tasks. While simple and interpretable, its performance can degrade in high-dimensional spaces or with noisy data.
- **Deep Learning Architectures**: Deep learning extends ANNs by introducing multiple hidden layers and non-linear activation functions. For example, convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been used with satellite imagery and time series data to model spatiotemporal variations in water quality.
- **Hybrid and Ensemble Models**: Recent studies have explored hybrid architectures that combine ML models (e.g., ANN-SVR or RF-ANN), as well as ensemble techniques such as stacking and voting, to improve robustness and generalization in water quality forecasting tasks.

Integrating these machine learning models into water quality monitoring systems offers significant advantages, including real-time analysis, automated anomaly detection, and predictive modeling under data-scarce conditions. This research applies such approaches to the Jacarepaguá Lagoon System, aiming to support evidence-based management and align with global sustainability objectives using intelligent, data-driven methods.

2.4 Jacarepaguá Lagoon System

The region of Jacarepaguá Lagoon System (JLS) is located in the West Zone of Rio de Janeiro, Brazil, encompassing rapidly urbanizing neighborhoods such as Barra da Tijuca, Jacarepaguá, and Rio das Pedras. Since the 1970s, this region has undergone substantial demographic expansion, with its population nearly tripling over the past few decades. The JLS consists of a complex network of interconnected coastal lagoons—Lagoa da Tijuca, Lagoa de Jacarepaguá, Lagoa do Camorim, and Lagoa de Marapendi—which are hydrologically connected and influenced by both freshwater inputs (e.g., rivers and stormwater runoff) and tidal exchanges with the Atlantic Ocean via the Joatinga Canal, as presented in Fig. 1. As the primary drainage basin for several densely populated urban areas, the JLS plays a vital ecological and infrastructural role by supporting diverse aquatic biodiversity and mitigating flood risks. However, the system faces significant environmental pressures due to accelerated urbanization, unregulated land occupation, and insufficient wastewater management. These factors have intensified water quality degradation, underscoring the importance of continuous environmental monitoring and sustainable urban planning to conserve this sensitive coastal ecosystem.

3 Methodology

The first practical stage of the research was to collect historical data on water quality monitoring in the JLS (Fig. 1), taken from INEA's indoor monitoring raw data spreadsheet, available on the agency's website [17–19].

We checked which parameters had the lowest rate of temporal discontinuity, resulting in 10 potential variables as presented in Table 1. Of these, 9 were selected as predictive variables (features), and the remaining (BDO) was determined as a target variable. In addition to the water quality data, maximum rainfall data was collected in the lagoon system region at five different time intervals: 15 min, 1 h, 4 h, 24 h, and 96 h. This data was obtained from the Sistema Alerta Rio website for the Barrinha station closest to the ocean. The rainfall data was merged with the water quality data using a standardized reference time. As all the data collected by INEA occurred after 5 a.m., this was the time selected to extract the rainfall data and append it to the other data.

Next, an initial analysis of the data involved the Cluster Analysis (CA) process using the k-means method. The aim was to establish groups containing the collection points based on the similarity between each parameter. Outliers were



Source: INEA (2013).

Fig. 1. Lagoons and data collect point in JLS.

identified and removed using Tukey's method. The data preprocessing procedure is detailed in the flowchart shown in Fig. 2.

For the modeling stage, initial models were selected based on those most frequently used in recent studies on computational modeling applied to coastal water quality assessment, along with experimental models developed in the context of this research. The selected models included Artificial Neural Networks (ANN) [9], Support Vector Regressor (SVR) [7], Decision Tree Regressor (DTR), and Random Forest Regressor (RFR). Considering the diversity of ANN and SVM specializations, these were subdivided into two and three variations, respectively. For neural networks, both a simple linear regression model and a deep learning model with multiple layers were created. For the SVM-based models, three kernels were used: Linear, Polynomial, and Radial Basis Function (RBF). Table 2 presents the list of models and their corresponding training hyperparameters.

Before training, the data was normalized using min-max scaling to avoid bias among variables. After rescaling, the dataset was split into training and testing sets, with 90% of the data used for training and 10% reserved for testing the model predictions and evaluating their performance. The metric used to assess model performance was the Mean Squared Error (MSE), computed as the mean of the squared prediction errors.

The training focused on the data point with the highest parameter dynamics, aiming to evaluate the models under the most complex scenario. Thus, Point 1—near the Joatinga Canal—was selected for model training.

Table 1. List of Features

n	Parameter	Abbreviation	Unity	Source
1	Dissolved Oxygen	DO	mg/L	INEA
2	Total Phosphorus	TF	mg/L	INEA
3	Turbidity	–	UNT	INEA
4	Conductivity	CE	(µS/cm)	INEA
5	Ortho-Phosphate Dissolved	OPD	mg/L	INEA
6	Coliforms	CT	NMP/100 mL	INEA
7	pH	–	–	INEA
8	Water Temperature	T	Celsius	INEA
9	Maximum Precipitation in (time range)	PM	mm	AlertaRio
10	Biochemical Oxygen Demand	BOD	mg/L	INEA

Each was trained over 50 generations using randomly sampled data to establish a comparative criterion among models. The Mean Square Error (MSE) was used to evaluate the methods presented in Eq. 1, and central tendency statistics were calculated. At the end of each generation, the model with the lowest MSE was assigned a point. After all generations were completed, a final ranking was produced based on the accumulated scores, listing the models in descending order of performance.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

where n is the size of the sample, in the units of observed (y_i) and predicted ($\hat{y}_i$) values.

4 Numerical Results

The analyses performed during the research were carried out using the Python 3.12 programming language, equipped with libraries for carrying out the necessary operations: *NumPy* for algebraic operations, *Scipy* and *Statistics* for tests and statistical operations, *Scikitlearn* and *Keras by Tensorflow* for training the models and numerical analysis.

Three groups were defined for the cluster analysis based on the elbow method. After model fitting, the total variance among centroids was 1256.92. Figure 3 presents the 3D scatter plot of the principal components (PCs), with data points colored according to their respective clusters.

Once the data points were assigned to clusters, a percentage distribution table of the sampling points was generated, showing results consistent with their spatial distribution.

The dendrogram in Fig. 4 illustrates a well-defined cluster separation. Points 1 and 2 (orange), closest to the ocean, form a distinct group with high inter-cluster Euclidean distance relative to the other points (green) but low distance

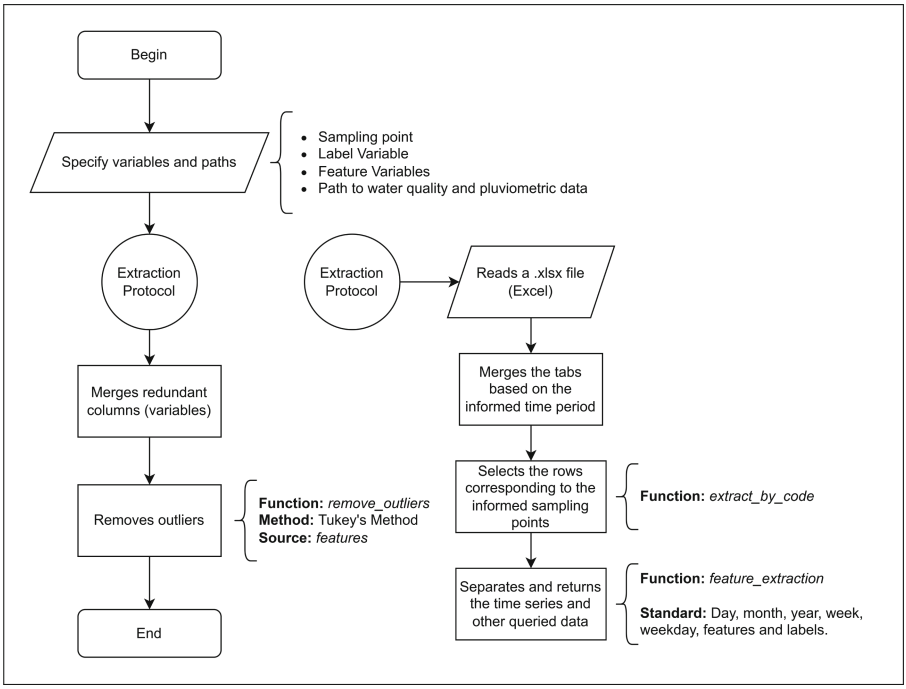


Fig. 2. Data preprocessing procedure.

between themselves. Points 6 and 3 were the most similar among the inland points, likely due to their proximity to river mouths and canals that discharge effluents into the system. Point 5, located in the Camorim Lagoon, showed moderate divergence from Points 6 and 3 in the same group, possibly explained by its morphology, which favors pollutant accumulation.

Point 4, in turn, formed an isolated third group due to its distinct behavior. Located in the central region of the lagoon system, it exhibited mixed characteristics—typical of ocean water and domestic sewage contamination—such as elevated conductivity. This singular combination of features led to its classification as an outlier in the clustering analysis.

Statistical validation through the Mann-Whitney U test for Points 1 and 2 yielded significant results, as did the Kruskal-Wallis test for Points 3, 5, and 6, showing p-values greater than 0.95, indicating strong internal consistency within these groups. No similarity test was applied to the third group, as it consisted solely of Point 4.

The seven models were trained over 50 generations using randomly sampled datasets. The performance analysis revealed clear trends, as summarized in Table 3.

SVR-based models showed consistent performance, particularly SVR-L ($\bar{x} \approx 17.2 \pm 11.4$) and SVR-P ($\bar{x} \approx 21.1 \pm 13.0$). In contrast, SVR-RBF underper-

Table 2. Models and training hyperparameters.

Name	Abv.	Architecture/Hyperparameters
Basic Neural Networks	BANN	Sequential model with 2 processing layers; Perceptrons: (16-1); Activations: (elu-selu); Optimizer: Adam
Deep Neural Networks	DANN	Sequential model with 2 processing layers; Perceptrons: (128-(Dropout 0.25)-64-16-1); Activations: (tanh-elu-selu-relu); Optimizer: Adam
Support Vector Regressor - Linear	SVR-L	Default configuration from the *sklearn* library
Support Vector Regressor - Polynomial	SVR-P	Default configuration from the *sklearn* library
Support Vector Regressor - RBF	SVR-RBF	Default configuration from the *sklearn* library
Decision Tree Regressor	DTR	Max Depth: Length of the training data vector; Min Samples Split: Total data vector length divided by 12
Random Forest Regressor	RFR	Number of Estimators: 64

Table 3. Model performance statistics over 50 iterations.

–		MSE			
Nome	*Score*	$\bar{x}$	std	Max	Min
DTR	15	17.0	12.6	54.6	3.6
SVR-L	14	17.2	11.4	54.9	1.8
RFR	7	29.0	23.2	92.3	1.6
SVR-P	7	21.1	13.0	55.2	1.7
DANN	4	29.0	18.2	90.0	0.8
SVR-RBF	3	42.9	32.7	196.9	7.6
BANN	0	70.3	38.4	168.8	7.8

formed, with the highest maximum MSE and the second largest standard deviation. The Decision Tree Regressor (DTR) achieved the best overall performance ($\bar{x} \approx 17.0 \pm 12.6$), with a score slightly higher than SVR-L. However, DTR also presented a higher minimum error, suggesting less consistency in lower-error scenarios.

Random Forest Regressor (RFR) ranked third ($\bar{x} \approx 29.0 \pm 23.2$), but its overall score was significantly lower than SVR-L. Neural network-based models exhibited suboptimal results. BANN performed the worst across all metrics and received no score. DANN matched RFR in mean MSE but demonstrated slightly lower variance ($\bar{x} \approx 29.0 \pm 18.2$). Figure 5 presents the MSE trajectory of each model across the 50 generations.

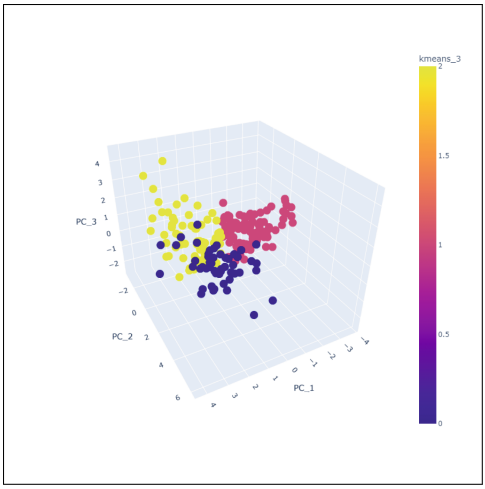


Fig. 3. 3D scatter plot of principal components and clusters.

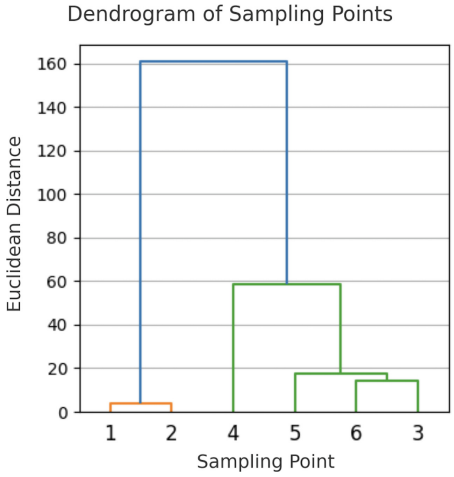


Fig. 4. Dendrogram of cluster group percentages.

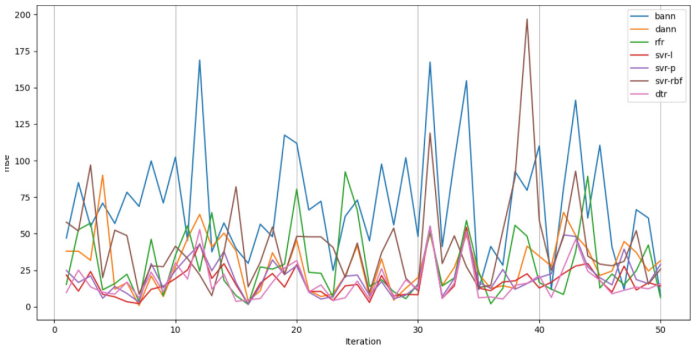


Fig. 5. Mean Squared Error (MSE) of models across 50 iterations.

5 Final Remarks

Monitoring water quality in estuarine environments is essential for sustainable urban development and aligns closely with the United Nations Sustainable Development Goals, particularly SDG 6 (Clean Water and Sanitation) and SDG 11 (Sustainable Cities and Communities). This study presented a preliminary approach to assessing water quality in the Jacarepaguá Lagoon System (JLS), using Biochemical Oxygen Demand (BOD) as the target variable.

Several machine learning models were evaluated, including Decision Tree Regressor (DTR), Random Forest Regressor (RFR), Support Vector Regressors with different kernels (SVR-L, SVR-P, SVR-RBF), and neural network-based models (DANN and BANN). Among them, the DTR model achieved the best performance, with a Mean Squared Error (MSE) of approximately 17 $(\text{mg/L})^2$.

While the initial results are encouraging, we acknowledge key limitations such as data scarcity, reliance on a single target variable, and the absence of spatial and temporal modeling. To address these issues, future work will incorporate remote sensing data, apply data augmentation techniques, and expand the analysis to include additional water quality indicators (e.g., turbidity and dissolved oxygen). We also plan to employ advanced models capable of capturing spatiotemporal dynamics, such as Long Short-Term Memory (LSTM) networks, and evaluate performance using more comprehensive metrics like F1-score and R^2.

These enhancements aim not only to improve predictive accuracy and model robustness but also to provide actionable insights that support evidence-based environmental management and policymaking.

Acknowledgement. The authors would like to thank the following Brazilian Agencies: CAPES and the National Council for Scientific and Technological Development - CNPq. DB thanks Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) for the financial support through grants E-26/210.798/2024. PHG thanks Federal University of Rio de Janeiro (23079.227622/2023-70), National Council for Scientific and Technological Development - CNPq (307663/2021-3) and

Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro - FAPERJ (E-26/201.341/2022) for the financial support.

References

1. Chacón Abarca, S., et al.: Understanding the dynamics of a coastal lagoon. Geosciences **11**(8), 301 (2021)
2. Silva, R., et al.: A framework to manage coastal squeeze. Sustainability **12**(24), 10610 (2020)
3. de Araújo, A., et al.: General system architecture and COTS prototyping of an AIoT-enabled sailboat. IEEE IoT J. **11**(3), 3801–3811 (2024)
4. Pinto, F.P., et al.: Seasonal assessment of water quality in Mirim Lagoon. Anais da ABC (2020)
5. Ferrão-Filho, A., Moscatelli, M.: Recreational risk associated with the presence of cyanobacteria (in Portuguese). SciELO Preprints (2020)
6. Preece, E.P., et al.: A review of microcystin detections. In: Harmful Algae, pp. 31–45 (2017)
7. Sun, Z., Fan, Y.: A combined water quality prediction model using Spark. In: Aqua (2022)
8. Carlsen, L., Bruggemann, R.: The 17 United Nations' sustainable development goals. Int. J. Sust. Dev. World Ecol. **29**(3), 219–229 (2022)
9. Zhou, J., et al.: Water quality prediction using IGRA and LSTM. Water **10**, 1148 (2018)
10. Telles, W.R., et al.: Modeling surface waterflood with MOHID. In: Mathematics, Computers and Environment, pp. 125–143 (2019)
11. de Andrade Costa, D., et al.: Water quality estimates using ML. J. Hydroinf. **26**(11), 2798–2814 (2024)
12. de Oliveira, M., et al.: Mathematical modelling in microplastic research. Environ. Rev. **33**, 1–18 (2025)
13. Asadollah, S., et al.: Dissolved organic carbon estimation in lakes. Water Res. **277**, 123350 (2025)
14. Banerjee, S., et al.: Sensitivity of parameters in modeling hydrological services. Model. Earth Syst. Environ. **11**(4), 1–16 (2025)
15. Costa, D., et al.: Water quality estimates using machine learning techniques in an experimental watershed. J. Hydroinf. **26**(11), 2798–2814 (2024). https://doi.org/10.2166/hydro.2024.132
16. Zhou, J., et al.: Water quality prediction method based on IGRA and LSTM. Water **10**(9) (2018)
17. Instituto Estadual do Ambiente (INEA): Consolidated Water Quality Report for the Jacarepaguá Lagoons (2022). (in Portuguese). https://www.inea.rj.gov.br/wpcontent/uploads/2022/05/Lagoas-JPA-Consolidado-2020-2021.pdf
18. Instituto Estadual do Ambiente (INEA): Bathing Water Quality Bulletin – RJ (2024). (in Portuguese). https://www.inea.rj.gov.br/wpcontent/uploads/2024/02/Zona-oeste-e-Zona-sul-15-02-24.pdf
19. Instituto Estadual do Ambiente (INEA): Monitoring of Inland Waters (2024). (in Portuguese). https://www.inea.rj.gov.br/ar-agua-e-solo/como-e-feito-o-monitoramento-das-aguas-interiores/
20. Lamparelli, M.C.:Lamparelli, M.C.: Trophic States in Water Bodies of the State of São Paulo (in Portuguese). Ph.D. thesis, University of São Paulo (2004). https://www.teses.usp.br/teses/disponiveis/41/41134/tde-20032006-075813/pt-br.php