Machine Learning-Based Simulation of Monthly Water Quality in the Santa Lucía Chico River Basin

Pedro Pertusso[1[0009−0008−3872−4086]], Martina Pou[2[0009−0000−6710−6908]], Federico Vilaseca[2[0000−0002−5471−6165]], Alberto Castro[1,3[0000−0002−9174−398X]] and Angela Gorgoglione[2[0000−0002−2476−2339]]

[1] Department of Computer Science, School of Engineering, Universidad de la República, Montevideo 11300, Uruguay
[2] Department of Fluid Mechanics and Environmental Engineering, School of Engineering, Universidad de la República, Montevideo 11300, Uruguay
[3] Department of Electrical Engineering, School of Engineering, Universidad de la República, Montevideo 11300, Uruguay

Abstract. This study aims to develop a data-driven tool for monthly water quality simulation using machine learning techniques. The study focuses on the upper basin of the Santa Lucía Chico River in Uruguay, utilizing data from two water quality monitoring stations (XSLH010 and XSLH020). The variables considered include dissolved oxygen (DO), temperature (T), total nitrogen (NT), and phosphate (PO_4^{3-}). The time series data were split into training (80%) and testing (20%) sets, with separate min-max normalization applied to ensure consistent scaling across variables. The prediction models were trained using Extra Trees Regressor (ET) and Histogram-based Gradient Boosting Regressor (HGB), evaluated with Mean Absolute Error (MAE) and Mean Squared Error (MSE). This resulted in four models trained per variable. Nash-Sutcliffe Efficiency (NSE) was also calculated for model performance evaluation. Optimal hyperparameters were identified using a 5-fold cross-validation process and optimized with Optuna. The input dataset integrates domain knowledge by incorporating spatial dependencies, spatial correlations, physical dependencies, and temporal variability. Additionally, SHapley Additive exPlanations (SHAP) values were used to refine model inputs by removing low-importance variables. The models operate at a monthly time step, allowing for the assessment of long-term water quality trends. The results were highly satisfactory, with NSE values exceeding 0.6 for all variables across both stations, except for PO_4^{3-} at XSLH010. These findings demonstrate the potential of machine learning models for water quality prediction and provide a valuable tool for improving water resource management. Future efforts will focus on refining the model, incorporating additional data sources, and extending its applicability to other basins.

Keywords: Water quality modeling, Machine learning, Monthly prediction, Hydroinformatics.

1 Introduction

Water quality management is essential for maintaining sustainable water resources, particularly in regions experiencing significant anthropogenic pressure [1]. Traditionally, water quality assessments rely on physical and process-based models, which simulate hydrological and biogeochemical processes. While effective, these models often require extensive input data, complex parameterization, and significant computational resources, making their implementation challenging, especially in data-scarce regions [2]. Recent advancements in machine learning (ML) offer a data-driven alternative, enabling predictive modeling of water quality by identifying complex, nonlinear relationships between environmental variables and leveraging historical datasets [3, 4].

In many regions worldwide, water bodies are under increasing pressure due to agricultural expansion, industrial activities, and urbanization. These factors contribute to the degradation of water quality by introducing pollutants such as nutrients, heavy metals, and organic matter into aquatic ecosystems [5]. Predictive models play a vital role in assessing and mitigating these impacts, enabling better decision-making for sustainable water management. ML-based approaches, in particular, have gained traction due to their ability to process large and complex datasets, integrate spatial and temporal dependencies, and improve forecasting accuracy compared to traditional models [6, 7].

Uruguay, like many other countries, faces growing concerns over water pollution. The Santa Lucía River basin, a critical source of drinking water, has been significantly affected by nutrient enrichment, leading to algal blooms and eutrophication [8, 9, 10]. Despite ongoing monitoring efforts, the development of robust, data-driven predictive models remains limited. Adapting ML techniques to the region's specific conditions is essential to enhance water quality management and ensure the long-term sustainability of aquatic ecosystems.

The main objective of this study is to develop and evaluate ML-based models to predict key water quality parameters in the Santa Lucía Chico River Basin at a monthly timescale. By integrating spatial correlations, temporal dependencies, and domain knowledge, the proposed models aim to improve forecasting accuracy and provide actionable insights for water resource management. The study specifically focuses on predicting dissolved oxygen (DO), water temperature (T), total nitrogen (TN), and phosphate (PO_4^{3-}), using data from two monitoring stations (XSLH010 and XSLH020).

The methodology involves preprocessing historical water quality data, training and optimizing ML models, and evaluating their performance against observed data. This study contributes to the growing body of research on ML applications in water quality prediction and provides a framework for data-driven decision-making in water resource management.

2 Materials and Methods

2.1 Study area and data availability

The Santa Lucía Chico River basin, located in Uruguay, plays a vital role in the regional water supply. This study focuses on two monitoring stations, XSLH010 and XSLH020, where water quality data—including DO, T, TN, and PO_4^{3-}—have been collected. These variables serve as the output for the models developed in this study (Figure 1). Additionally, other water quality parameters, such as total phosphorus (TP), nitrate (NO_2^-), nitrite (NO_3^-), total solids (TS), turbidity (Turb) and conductivity (Cond), were recorded and used as model inputs.

These stations were selected for their strategic location in the upper basin, enabling an assessment of water quality trends before the river reaches the primary drinking water source for Montevideo and its surrounding areas [11]. Furthermore, hydrometeorological variables—including streamflow, precipitation, air temperature, solar radiation, heliophany, relative humidity, wind speed, and evapotranspiration—were incorporated as model inputs. Data for these variables were obtained from the hydrometric monitoring station in Florida, the meteorological station at INIA Las Brujas, and additional pluviometers (Cerro Colorado, La Cruz, San Gabriel, Sarandí Grande) (Figure 1).

Given the presence of significant missing data, this multivariate dataset was properly imputed at a monthly frequency before being used in this study [12].

The basin has experienced significant water quality degradation due to increasing agricultural activities and urban expansion. Elevated nutrient levels, particularly nitrogen and phosphorus, have contributed to periodic algal blooms and eutrophication events, underscoring the need for improved predictive models to support effective water resource management.

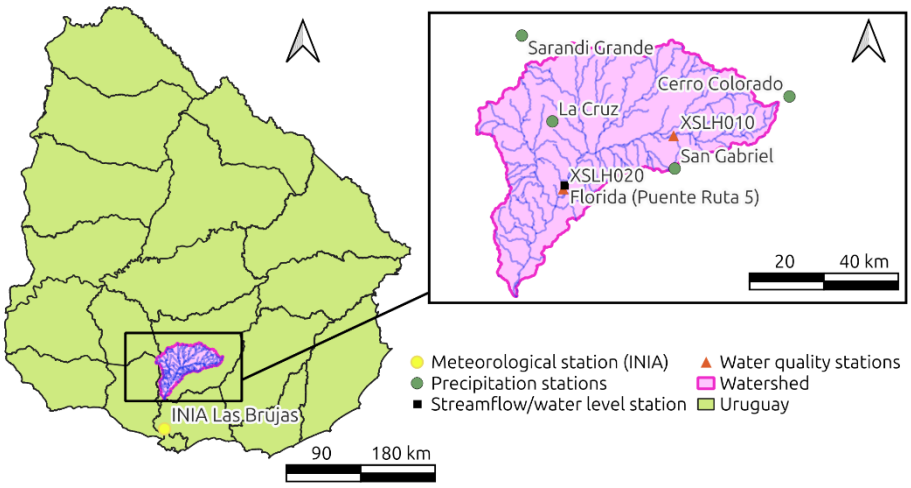

Fig. 1. Study area and location of the monitoring stations.

2.2 Data pre-processing and modeling optimization

The modeling process began by dividing the original dataset into two subsets: 80% for training and 20% for testing. Prior to model development, all variables were normalized using min-max scaling to ensure consistency in scale and to facilitate algorithm convergence. To enhance model accuracy, the dataset was further enriched with key hydrometeorological variables, including daily precipitation, mean air temperature, and surface runoff.

Two machine learning algorithms were selected for comparative analysis: the Extra Trees Regressor (ET) and the Histogram-based Gradient Boosting Regressor (HGB). These ensemble methods were chosen for their robustness in handling nonlinear relationships and multivariate datasets. Model performance was primarily assessed using Mean Absolute Error (MAE) and Mean Squared Error (MSE). Additionally, the Nash-Sutcliffe Efficiency (NSE) coefficient was calculated to provide a domain-specific metric for evaluating the predictive skill of the models in hydrological contexts.

For each target variable, four models were trained (two algorithms × two metrics), and optimal hyperparameters were identified using a 5-fold cross-validation strategy. The hyperparameter optimization process was carried out using the Optuna framework, which employs Bayesian optimization techniques to efficiently search the parameter space.

To incorporate domain knowledge and reduce input redundancy, feature selection was guided by correlation analysis. Pearson, Spearman, and Kendall correlation coefficients were computed between each predictor and the target variable. Predictors with a median correlation magnitude below 0.5 ($|\rho| < 0.5$) across the three methods were discarded to retain only the most relevant features.

To address spatial dependencies, downstream stations were systematically excluded from the training set when predicting upstream variables. This strategy was implemented to avoid information leakage that could arise from including future (downstream) data in the training process. In parallel, physical dependencies between variables were respected; for instance, air temperature was explicitly included as an input when predicting water temperature, reflecting its known influence in the physical system.

Temporal dependencies in the data were refined using the Exponentially Weighted Moving Average (EWMA) technique (Eq. 1). This method assigns greater weight to recent observations, enhancing the model's responsiveness to short-term fluctuations while still retaining long-term trends.

$$\begin{cases} EWMA\,(Y_n) = \dfrac{\sum_{i=0}^{t}(1-\alpha)^i Y_{n-1-i}}{\sum_{i=0}^{t}(1-\alpha)^i} \\ \alpha = \dfrac{2}{t+1} \end{cases} \quad \text{(Eq. 1)}$$

where, Y represents the time series, n denotes the total number of observations, and α corresponds to the weighting factor. The parameter t defines the temporal window

applied during imputation, which varies depending on the type of variable: approximately two months for water quality data, one week for hydrometric measurements, and one day for climatic variables.

Finally, additional physical constraints integrated into the models, such as non-negativity of certain outputs and hydrologically plausible limits, are described in further detail in reference [12].

3 Results and Discussion

The machine learning (ML) models demonstrated strong predictive performance overall, with NSE values exceeding 0.6 for all water quality variables, except for PO_4^{3-} at station XSLH010 (Table 1). These results underscore the effectiveness of the modeling framework, particularly the inclusion of spatial and temporal dependencies and the integration of domain-specific knowledge during feature selection and preprocessing.

Among the variables, DO exhibited the most robust predictive accuracy, with NSE values greater than 0.85 at both monitoring stations. This performance highlights the models' ability to capture the temporal dynamics and diurnal variations typically associated with DO fluctuations, which are strongly influenced by temperature, flow conditions, and biological activity. The ET model outperformed the HGB at XSLH020, while the reverse was true at XSLH010. These differences suggest that local hydro-environmental conditions and data distributions play a significant role in determining which algorithm generalizes better at a given site.

T was the variable with the highest accuracy during training, achieving NSE values close to 0.99. However, this came with a noticeable drop in test performance (0.84 at XSLH020 and 0.77 at XSLH010) indicating a degree of overfitting, especially at the latter station. This discrepancy could stem from limited variability in temperature patterns during training or from changes in local environmental conditions not captured in the training period.

TN predictions were satisfactory, with NSE values of 0.80 at XSLH020 and 0.60 at XSLH010. The lower performance at XSLH010 may reflect greater spatial heterogeneity in nitrogen sources, such as diffuse runoff or intermittent discharges, which were not fully captured by the selected input features. This result suggests that additional explanatory variables (e.g., land use, upstream agricultural practices) or finer spatial resolution could improve model performance.

PO_4^{3-} proved to be the most challenging variable to model, particularly at XSLH010, where the best-performing model achieved only 0.28 NSE during training and 0.26 during testing. Such low predictive skill indicates that key drivers of phosphate variability, such as episodic releases, sediment interactions, or localized anthropogenic inputs, may not be adequately represented in the current input dataset. This finding suggests the need for either richer input data (e.g., point-source locations, in-stream processes) or alternative modeling strategies, such as hybrid models combining process-based and data-driven approaches. In contrast, PO_4^{3-} modeling at XSLH020 was markedly better (0.81 NSE for the test set), reinforcing the influence of site-specific characteristics in determining model reliability and the importance of local calibration.

Overall, the results confirm the value of tailoring ML models to local conditions, the importance of feature engineering, and the potential limitations of purely data-driven approaches when key environmental drivers are unobserved or poorly quantified.

Table 1. Summary of the model's performance.

Variable	Station	Best model	Train - NSE	Test - NSE
DO	XSLH020	ET (MSE)	0.76	**0.88**
	XSLH010	HGB (MAE)	0.88	**0.85**
T	XSLH020	ET (MAE)	0.99	**0.84**
	XSLH010	HGB (MAE)	0.90	**0.77**
TN	XSLH020	ET (MAE)	0.92	**0.80**
	XSLH010	ET (MAE)	0.79	**0.60**
PO_4^{3-}	XSLH020	ET (MSE)	0.99	**0.81**
	XSLH010	HGB (MSE)	0.28	**0.26**

4 Conclusions

This study demonstrates the potential of ML models for monthly water quality prediction in the Santa Lucía Chico River Basin. The models performed well for DO and T, achieving NSE values above 0.85 for DO and highlighting strong predictive capability. However, T models exhibited some overfitting, indicating the need for further refinement.

TN predictions were satisfactory, with better accuracy at XSLH020 than XSLH010, suggesting the influence of local environmental factors. PO_4^{3-} predictions were the least accurate, particularly at XSLH010, where performance remained below 0.30 NSE, emphasizing the need for additional predictors or alternative modeling approaches.

To improve model robustness, physical constraints were introduced in the model implementation. This approach helped enhance interpretability and reduce information leakage, though further refinements are needed to optimize its impact.

Overall, while the models effectively captured temporal water quality trends, improvements in feature selection, additional explanatory variables, and strategies to enhance generalization could further refine their accuracy. Future work should focus on addressing overfitting, incorporating more spatial and environmental factors, and validating the models in other watersheds to assess their broader applicability.

Acknowledgments. This work was supported by the National Research and Innovation Agency (ANII) [grant numbers FMV-3-2022-1-172720].

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Pou, M., Pastorini, M., Alonso, J., Gorgoglione, A.: Exploring the nexus between water quality and land use/land cover change in an urban watershed in Uruguay: a machine learning approach. Environ Sci Pollut Res, 31, 48687–48705 (2024)
2. Boratto, T.H., Campos, D.E., Fonseca, D.L., Soares Filho, W.A., Yaseen, Z.M., Gorgoglione, A., Goliatt, L.: Hybridized machine learning models for phosphate pollution modeling in water systems for multiple uses. Journal of Water Process Engineering 64, 105598 (2024)
3. Russo, C., Castro, A., Gioia, A., Iacobellis, V., Gorgoglione, A.: Improving the sediment and nutrient first-flush prediction and ranking its influencing factors: An integrated machine-learning framework. Journal of Hydrology 616, 128842 (2023)
4. Gorgoglione, A.; Gioia, A.; Iacobellis, V. A Framework for assessing modeling performance and effects of rainfall-catchment-drainage characteristics on nutrient urban runoff in poorly gauged watersheds. Sustainability 11, 4933 (2019)
5. Gorgoglione, A., Castro, A., Iacobellis, V., Gioia, A.: A comparison of linear and non-linear machine learning techniques (pca and som) for characterizing urban nutrient runoff. Sustainability 13 (2021)
6. Rodríguez, R., Pastorini, M., Etcheverry, L., Chreties, C., Fossati, M., Castro, A., Gorgoglione, A.: Water-Quality Data Imputation with a High Percentage of Missing Values: A Machine Learning Approach. Sustainability 13, 6318 (2021)
7. Cal A, Pastorini M, Tiscornia G, Rivas-Rivera N, Gorgoglione A.: Assessing dependence between land use/land cover and water quality: A comparison at a small and a large watershed in Uruguay. Agrocienc Urug. 27(NE1):e1192 (2023)
8. Gorgoglione, A., Gregorio, J., Ríos, A., Alonso, J., Chreties, C., Fossati, M.: Influence of Land Use/Land Cover on Surface-Water Quality of Santa Lucía River, Uruguay. Sustainability 12, 4692 (2020)
9. Aubriot, L., Delbene, L., Haakonson, S., Somma, A., Hirsch, F., Bonilla, S.: Evolución de la eutrofización en el Río Santa Lucía: Influencia de la intensificación productiva y perspectivas. Innotec 14, 7–17 (2017)
10. Goyenola, G., Meerhoff, M., Teixeira-de Mello, F., González-Bergonzoni, I., Graeber, D., Fosalba, C., Vidal, N., Mazzeo, N., Ovesen, N.B., Jeppesen, E., et al.: Phosphorus dynamics in lowland streams as a response to climatic, hydrological and agricultural land use gradients. Hydrol. Earth Syst. Sci. Discuss. 12, 3349–3390 (2015)
11. Vilaseca, F., Castro, A., Chreties, C., Gorgoglione, A.: Assessing influential rainfall-runoff variables to simulate daily streamflow using random forest. Hydrol Sci J, 68(12), 1738-–1753 (2023)
12. Pastorini, M., Rodríguez, R., Etcheverry, L., Castro, A., Gorgoglione, A.: Enhancing environmental data imputation: A physically-constrained machine learning framework. Science of the Total Environment 926, 171773 (2024)