Journal of Posthumanism
2025
Volume: 5, No: 6, pp. 2276–2291
ISSN: 2634-3576 (Print) | ISSN 2634-3584 (Online)
posthumanism.co.uk

DOI: https://doi.org/10.63332/joph.v5i6.2335

Advanced Machine Learning Techniques for Hydrological Prediction of Lake Titicaca

Lenin Huayta-Flores[1], Hugo Yosef Gomez-Quispe[2], Robert Antonio Romero-Flores[3]

Abstract

Global climate change is contributing to significant water reduction, evident in Lake Titicaca, where water levels have declined markedly in recent years. This study employs ten advanced machine learning algorithms to improve the prediction accuracy of water level behavior in the Lake Titicaca basin. The models utilize year, month, day, precipitation, maximum temperature, and minimum temperature as input variables. Accurate predictions are crucial for informing multidisciplinary planning related to agricultural and livestock production and for developing contingency strategies against droughts and floods. Historical data were sourced from the SENAMHI ENAFER pier station in Puno. Comparative evaluation indicates that the Gradient Boosting Regressor algorithm achieves superior performance based on multiple metrics: MAE = 0.00035279, MSE = 0.00000247, RMSE = 0.0015725, Coefficient of Determination R^2 = 0.99999706, and Cross-Validation (average RMSE = 0.00199047). This research underscores the potential of machine learning algorithms in hydrology, offering valuable insights applicable to basin modeling worldwide.

Keywords: *Machine Learning, Water Resources Management, Lake Titicaca, Lake Level, Hydrological Prediction.*

Introduction

Improving the accuracy of predicting Lake Titicaca's hydrological dynamics is essential for the sustainable management of its resources, particularly as the region heavily depends on the lake's water balance. Traditional modeling approaches, including conventional physical and statistical models, face limitations in capturing the complex interplay of climatic and anthropogenic factors influencing the lake (Kundzewicz et al., 2014). These factors encompass climate change, interannual precipitation variability, glacier melt, and human activities such as intensive agriculture and urbanization. Advanced machine learning methods offer a promising alternative, capable of processing large, heterogeneous datasets, identifying non-linear patterns, and enhancing predictive capabilities (Reichstein et al., 2019). The application of these tools supports regional water resource planning by facilitating informed decisions on water allocation and conservation, while also advancing scientific understanding within the field of hydrology (Shen et al., 2018).

Lakes serve critical roles in water storage for consumption, hydroelectric power generation, and various environmental, agricultural, and industrial purposes. Optimizing lake utilization necessitates accurate lake water level (LWL) prediction, a key challenge in water resources

[1] Universidad Nacional del Altiplano, Puno, Peru, Email: lhuayta@unap.edu.pe, ORCID: https://orcid.org/0000-0003-2759-9095 (Corresponding Author).
[2] Universidad Nacional del Altiplano, Puno, Peru. E-mail: hygomez@unap.edu.pe, ORCID: https://orcid.org/0000-0002-8627-412X
[3] Universidad Nacional del Altiplano, Puno, Peru, Email: rromero@unap.edu.pe, ORCID: https://orcid.org/0000-0002-6144-9309.

management (Mohammadi et al., 2020). Fundamentally, hydrology seeks to understand the hydrological cycle, including its surface and subsurface processes and interactions. This understanding necessitates precise measurements. While measurement techniques for flow and precipitation have advanced, hydrological models remain indispensable tools for comprehending the behavior of the cycle's components. Such models are vital for effective water management and planning (Cabrera, 2012). The present study focuses on characterizing the water potential within the Huancané, Ramis, Coata, and Ilave river basins, which are part of the Titicaca hydrographic region (Metzger Terrazas, 2017). Lake water level fluctuation is inherently complex and dynamic, exhibiting significant stochasticity and non-linearity, making it challenging to model and forecast accurately (Ozdemir et al., 2023; Sannasi Chakravarthy et al., 2022; Wang & Wang, 2020; Zhu, Lu, et al., 2020).

Lake water level is a key physical indicator of lacustrine ecosystem health; its fluctuations significantly impact these ecosystems, making accurate prediction crucial for effective management (Zhu, Hrnjica, et al., 2020). Lakes are also vital reservoirs for drinking water, hydropower, and other industrial and agricultural uses (Cabrera, 2012; Mohammadi et al., 2020). Understanding the water cycle requires continuous, precise monitoring of hydrological data for efficient water resource management, disaster mitigation, and economic planning (Huang et al., 2021). Integrating data from in-situ measurements and remote sensing into machine learning models enhances hydrological predictions, thereby improving our understanding of lake dynamics and water management strategies (Huang et al., 2021; Zhu, Lu, et al., 2020). Supported by hydrological models, machine learning algorithms like multiple linear regression, random forest, and k-nearest neighbors have significantly improved the prediction of water level fluctuations, particularly when combined with historical and satellite data (Kumar et al., 2023; Ozdemir et al., 2023; Singh et al., 2024; Wang & Wang, 2020). Accurate water level prediction is essential for managing water resources, preventing floods, and optimizing lake usage, tasks increasingly facilitated by artificial intelligence and satellite data utilization (Deng et al., 2022; Tan et al., 2023).

Since 2014, Lake Titicaca has shown a continuous decrease in its average annual level. However, the 2021-2022 hydrological year recorded levels exceeding those of the preceding six years (SENAMHI, 2024). This study aims to develop and apply advanced machine learning techniques to enhance the accuracy and predictive power of hydrological models used for assessing Lake Titicaca's behavior in the Puno region. We employ several machine learning models (Cutipa et al., 2020; Sulca et al., 2024; Sulca Jota et al., 2024), including Multiple Linear Regression, Ridge, Lasso, ElasticNet, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor, Hist Gradient Boosting Regressor, Extra Trees Regressor, and K-Nearest Neighbors Regressor. These models are applied to improve Lake Titicaca water level predictions using daily data from the 1982-2012 period. Notably, the 1943-1944 hydrological year represents the most significant deficit in the historical record, with an anomaly of -2.72 m relative to the average level of 3809.45 m (SENAMHI, 2024).

The remainder of this paper is organized as follows: Section 2 details the "*materials and methods"* used in the study. Section 3 presents the main "*results"*, while Section 4 explores "*alternative techniques"* for hydrological prediction. Finally, Section 5 summarizes the key "*conclusions"* and implications of the findings.

Materials and Methods

Study Area

Lake Titicaca is situated in southeastern Peru and northwestern Bolivia (Figure 1), spanning coordinates 68° 33′ – 70° 1′ W and 15° 6′ – 16° 50′ S. The lake comprises two main sections: Lago Mayor (Large Lake) to the north and Lago Menor (Small Lake) to the south. Puno Bay, the inner bay, lies on the western side, bordered by the Capachica and Chucuito peninsulas. The outer bay contains the Titicaca National Reserve and the Huile River, a tributary (Biamont-Rojas, 2022). As an essential water resource, the lake plays a vital role in agriculture, supports local fauna, flora, and the lakeside population (circum-lacustrine), and contributes to ecological balance. It also significantly influences local household economies. Nevertheless, the lake has been experiencing a considerable annual decrease in water volume. A recent study (1992-2020) using satellite imagery estimated an annual water loss of approximately 120 million metric tons (Lujano et al., 2023).

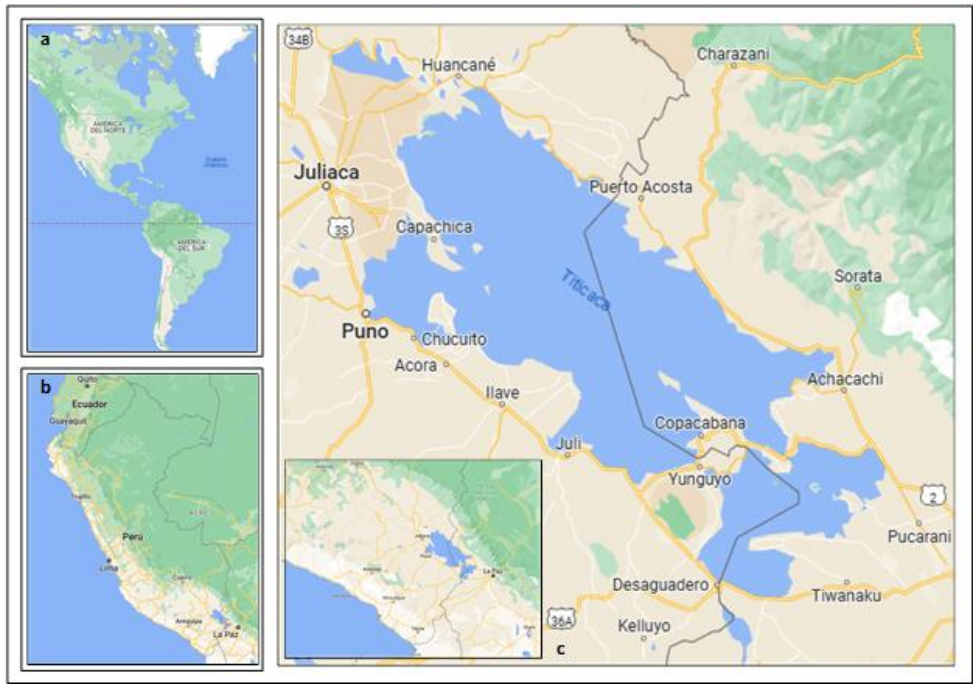

Figure 1. Location of the study area. (a) Peru in South America, (b) Lake Titicaca in Peru, (c) Puno Bay (ENAFER Station) in Lake Titicaca, southern Peru

Data Used

Two datasets were obtained from the National Service of Meteorology and Hydrology (SENAMHI) data repository. The first, a flat file (qc00000708.txt), provides detailed daily meteorological data (year, month, day, precipitation, maximum temperature, minimum temperature) from February 1, 1964, to December 31, 2012. These data were merged with monthly records of the Lake Titicaca water level (September 1, 1982, to December 31, 2024), downloaded from the SENAMHI monitoring portal, the website address is https://www.senamhi.gob.pe/main.php?p=monitoreo-informacion-mensual (accessed on December 2024). This combined information is crucial for analyzing climate patterns and

variability (Picado, 2017) that directly and indirectly impact water resource vulnerability, management, and availability.

The meteorological data analysis is based on a total of 17,687 records collected over a substantial time span from 1964 to 2012. The dataset indicates that average daily precipitation levels are relatively low, with a mean value of 2.01 mm, reflecting the region's generally dry climatic conditions. Maximum daily temperatures during the period range from 3.00°C to 22.80°C, with an overall mean of 15.08°C, suggesting moderate thermal conditions during daytime. In contrast, minimum temperatures show a broader variation, ranging from -7.20°C to 10.40°C, and an average value of 2.74°C, pointing to significantly cooler conditions during nighttime, including occurrences of subzero temperatures. Overall, the long-term meteorological records reveal temporal stability, characterized by low variability in precipitation and a consistent pattern of moderate thermal conditions. These findings provide a foundational understanding of the region's climate, which appears to be relatively stable over the nearly five-decade period under study.

The evolution of average temperatures (°C) over the four decades of the 1980s, 1990s, 2000s, and 2010s is shown in Figure 2. The data reveal a fluctuating yet generally upward trend in both maximum and minimum temperature values throughout the observed period. Notably, maximum temperatures reached a pronounced peak around 1997, registering approximately 18°C, which may reflect a particularly warm climatic episode during that time. Minimum temperatures, in contrast, remained relatively stable during the 1980s and 1990s, fluctuating between 2°C and 4°C, before exhibiting a gradual increase in the subsequent decades. The 2010s are characterized by elevated average maximum temperatures, approaching 17.5°C, which reinforces the indication of a sustained warming trend. The observed decadal distinctions in temperature behavior suggest the presence of broader climatic shifts or underlying seasonal anomalies that may be influencing regional thermal dynamics. These patterns underscore the importance of continued monitoring and long-term climatic analysis to better understand potential impacts associated with global or localized climate change.

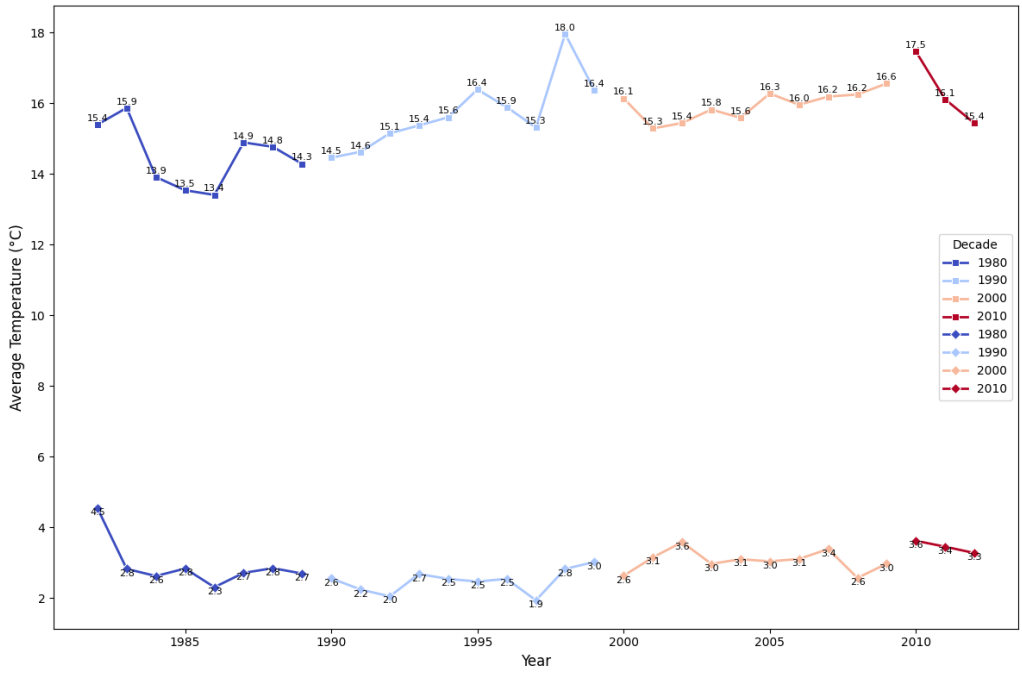

Figure 2. Average Maximum and Minimum Temperatures by Year

Figure 3 presents the average annual precipitation data for the period 1980–2010, grouped by decade to facilitate the analysis of long-term trends. During the 1980s, greater variability in precipitation levels is observed, with values ranging between 1.0 and 3.6 mm, suggesting a phase of climatic instability or more marked interannual fluctuations. In contrast, the following decades show a notable reduction in this variability, with averages remaining more consistently within the range of 1.5 to 2.9 mm. From the 2000s onward, the data indicate a trend toward the stabilization of average precipitation levels, accompanied by smaller fluctuations compared to previous decades. Although the records corresponding to the 2010s are limited, a slight increase in average values is observed relative to the preceding decades. This behavior may indicate a transition toward a more stable precipitation regime in terms of annual averages; however, additional data are needed to confirm this possible trend with greater statistical certainty.

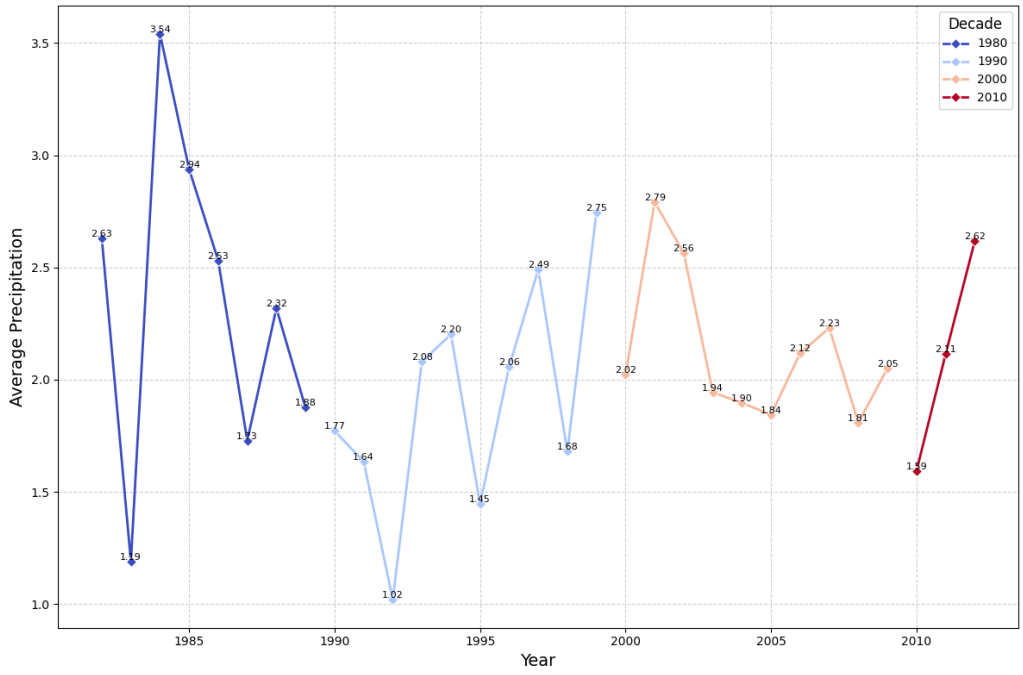

Figure 3. Average Precipitation by Year

The average monthly lake level measurements from 1982 to 2024, with each year represented individually to highlight interannual and seasonal dynamics, are visualized in Figure 4. The data reveal a pronounced and consistent annual cycle, characterized by peak water levels typically occurring between May and June, and troughs observed around November and December. Despite this seasonal regularity, a gradual and persistent decline in both peak and minimum lake levels is evident over the decades, indicating a long-term downward trend. Specifically, during the 1980s and 1990s, maximum lake levels frequently exceeded 3,811 meters above sea level. In contrast, more recent observations from the 2010–2024 period show consistently lower values, with peak levels barely reaching 3,810 meters and minimum levels falling below 3,809 meters. This pattern suggests a sustained reduction in overall lake water volume. Possible contributing factors to this decline include long-term effects of global climate change, diminished regional precipitation possibly associated with prolonged drought conditions, and increased anthropogenic water abstraction for agricultural, domestic, or industrial purposes. These findings underscore the need for integrated water resource management strategies and continued monitoring to mitigate potential ecological and socio-economic impacts.

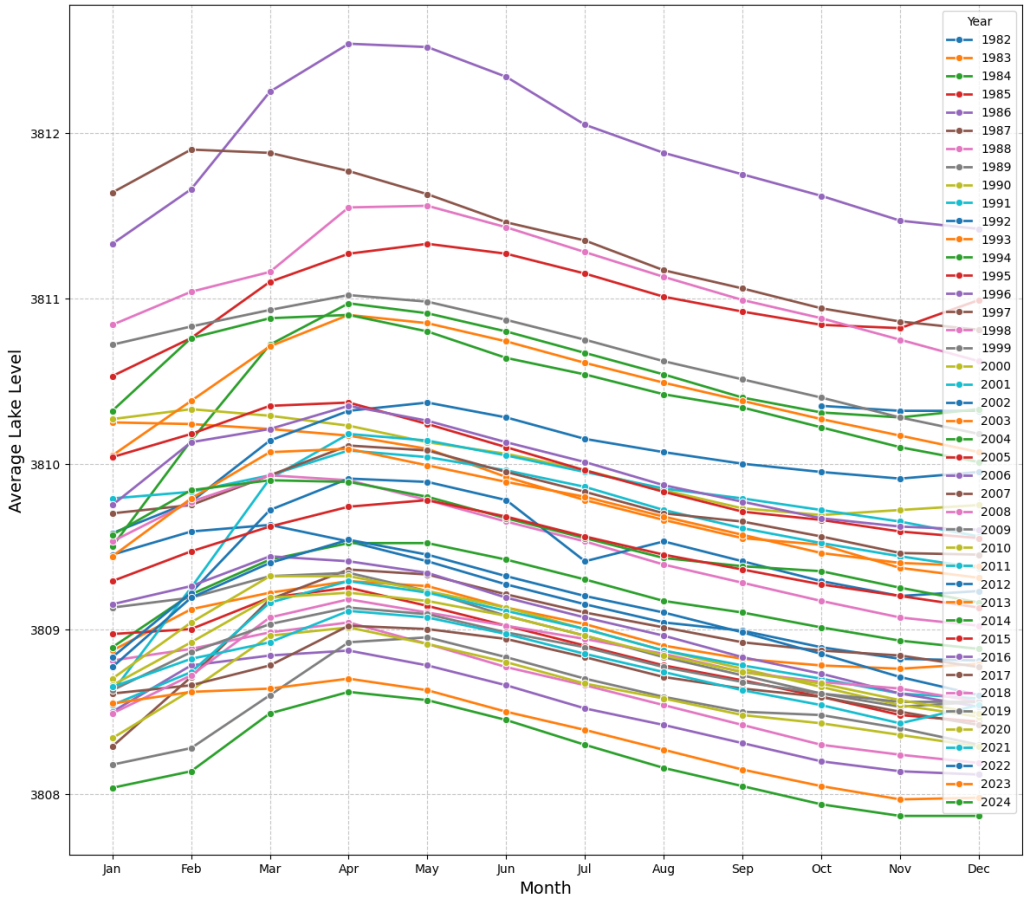

Figure 4. Average Lake Level by Month and Year

Figure 5 illustrates the decadal evolution of the average annual water level of Lake Titicaca, spanning from the 1980s through the 2020s, and reveals a pattern of notable fluctuations over time. During the 1980s, the lake experienced a sustained increase in water level, culminating in a historical maximum around 1986. This peak was followed by a marked and rapid decline. The 1990s were characterized by a more pronounced downward trend, with the lake reaching some of the lowest average levels recorded during the period under study. In the 2000s, a partial recovery was observed, with levels peaking near the year 2000 before initiating another phase of decline. The 2010s demonstrated a period of relative hydrological stability, with only minor oscillations in average annual levels. However, the early 2020s indicate a renewed decrease, approaching levels comparable to the historic lows seen in the 1990s. These fluctuations suggest a cyclical behavior in lake level dynamics, likely influenced by a combination of climatic variability—such as changes in precipitation patterns and temperature—and anthropogenic pressures, including water extraction and land-use changes. The observed trends highlight the critical importance of sustained hydrometeorological monitoring and integrated watershed management to ensure the long-term sustainability of Lake Titicaca as a vital regional resource.

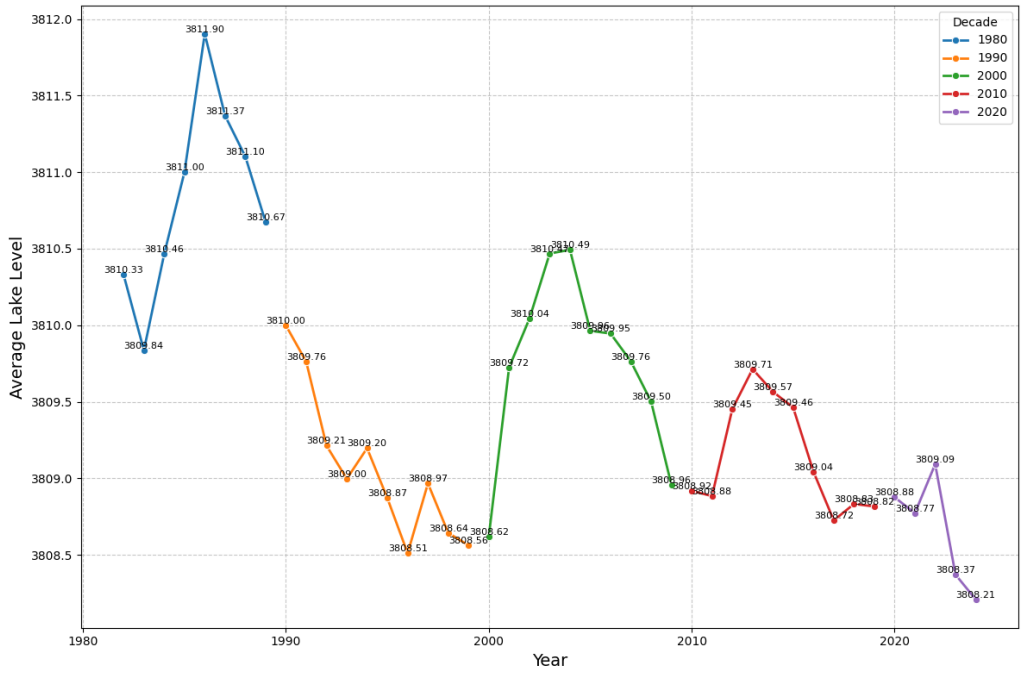

Figure 5. Average lake level by year

Figure 6 presents a heatmap displaying the correlation matrix among key variables, including year, month, day, precipitation, maximum temperature (Tmax), minimum temperature (Tmin), and lake level. The analysis reveals a moderate negative correlation between lake level and year ($r = –0.45$), as well as between lake level and Tmax ($r = –0.42$). These values suggest a discernible long-term decline in lake levels over the years, which appears to be associated with increasing maximum temperatures—a potential indicator of climate change impacts in the region. In contrast, the correlations between lake level and both precipitation ($r = –0.01$) and Tmin ($r = –0.07$) are notably weak, indicating that these variables have limited direct influence on lake level variations within the dataset. The relatively stronger association with Tmax, coupled with the declining trend over time, underscores the significance of thermal stress and evaporative processes as key drivers of hydrological changes in Lake Titicaca. These findings point to the potential predominance of temperature-related factors over precipitation in shaping the lake's dynamics and reinforce the urgency of climate-adaptive water resource management strategies.

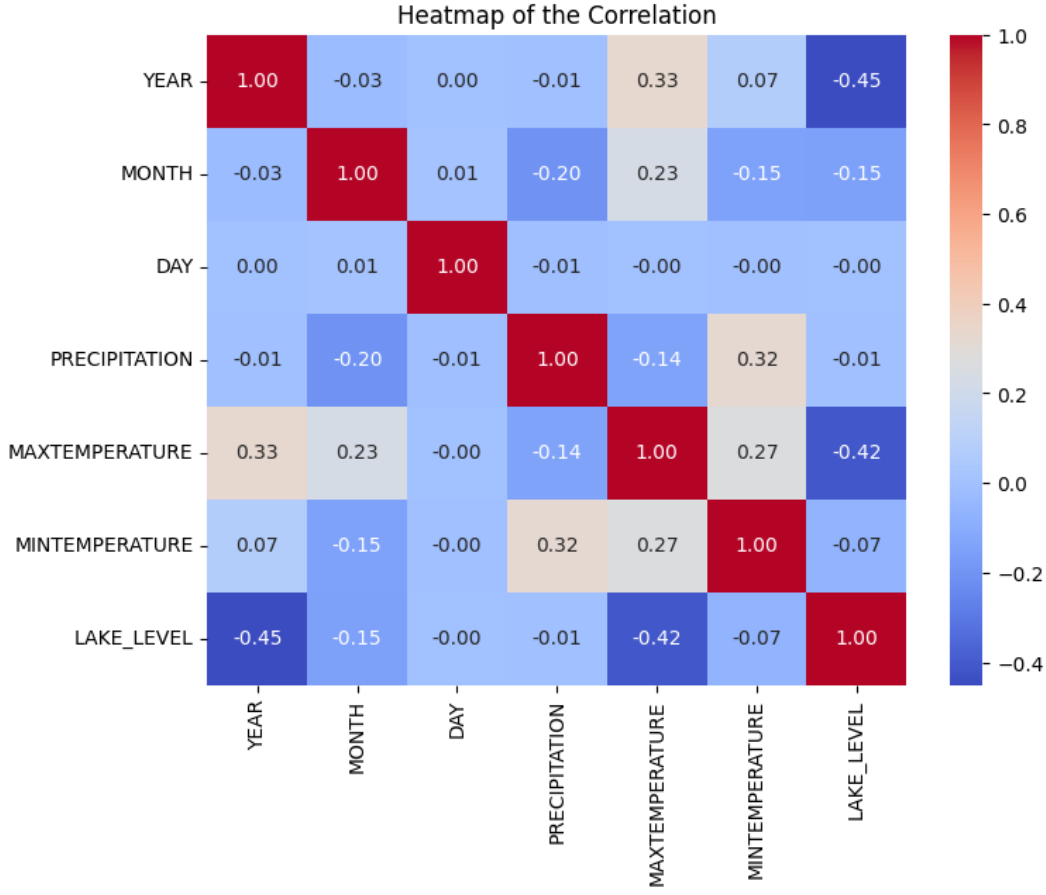

Figure 6. Heatmap of Correlations Between Variables

Machine Learning Algorithms

A variety of regression algorithms are commonly employed in predictive modeling tasks, each offering distinct methodological advantages depending on the nature of the dataset and the problem context. Traditional approaches such as Multiple Linear Regression serve as a foundational method, assuming linear relationships among variables. Regularization techniques like Ridge Regression, Lasso Regression, and ElasticNet enhance linear models by addressing issues of multicollinearity and overfitting, particularly in high-dimensional datasets. In contrast, Decision Tree Regression provides a non-parametric alternative capable of capturing nonlinear interactions without requiring extensive preprocessing. Ensemble-based methods such as Random Forests combine multiple decision trees to improve predictive accuracy and generalization through bagging techniques. Similarly, Boosting algorithms, particularly Gradient Boosting Regression, iteratively refine weak learners to build a strong predictive model, often achieving state-of-the-art performance in various regression tasks. The selection of an appropriate algorithm must be informed by data characteristics—such as linearity, noise, dimensionality—and the interpretability and performance requirements of the specific application.

Multiple Linear Regression (MLR) is a common statistical and machine learning algorithm used

to model a linear relationship between one or more predictors and a response variable. MLR extends simple linear regression by using multiple independent variables to predict a single dependent variable (Maulud & Abdulazeez, 2020), as shown in Equation 1.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...\beta_n x_n + \in \quad (1)$$

Ridge Regression is a type of regularized linear regression used for analyzing multicollinear data (highly correlated predictors). It penalizes the model's coefficients using an L2 norm to prevent overfitting. The objective function is similar to linear regression but includes an L2 penalty term, as shown in Equation 2, which is minimized during training.

$$J(\beta) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{n} \beta_j^2 \quad (2)$$

Lasso Regression (Least Absolute Shrinkage and Selection Operator Regression) is another regularized linear regression method. Unlike Ridge, it penalizes the sum of the absolute values of the coefficients (L1 norm). This L1 penalty can shrink some coefficients exactly to zero, effectively performing variable selection. The objective function includes the L1 penalty term (Equation 3).

$$J(\beta) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{n} |\beta_j| \quad (3)$$

Elastic Net Regression combines L1 (Lasso) and L2 (Ridge) penalties, offering a balance between the two. It can handle multicollinearity while also performing variable selection. The estimate can be viewed as a Bayesian posterior mode under a specific prior distribution (Hans, 2011). Its cost function incorporates both L1 and L2 terms (Equation 4).

$$J(\beta) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \lambda \left(\alpha \sum_{j=1}^{n} |\beta_j| + \frac{1-\alpha}{2} \sum_{j=1}^{n} \beta_j^2 \right) \quad (4)$$

Decision Trees and Ensembles are machine learning techniques used for classification and regression tasks. A decision tree is a model that represents decisions and their possible consequences through a tree-like structure, where each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf represents a prediction or class. Ensemble methods, on the other hand, combine multiple models (often decision trees) to improve overall performance; among the most common are Random Forest, which averages or votes on multiple randomly constructed trees, and Gradient Boosting, which sequentially builds trees by correcting the errors of the previous ones. These techniques are often robust, accurate, and effective for handling complex and nonlinear data. Decision Tree Regressor are popular algorithms known for their interpretability, as their structure resembles human decision-making processes (Pathak et al., 2018). Building a regression tree involves recursively splitting the data into nodes based on feature thresholds that minimize variance within nodes, forming decision rules. Predictions are typically the average target value of the instances within a leaf node (Equation 5).

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \quad (5)$$

Random Forest Regressor is an ensemble method constructing multiple decision trees during training. Each tree is trained on a bootstrapped sample of the data and considers only a random subset of features at each split. This introduces randomness, reducing variance and improving generalization compared to a single tree (Breiman, 2001).The final prediction is usually the average of the individual tree predictions. Gradient Boosting Regressor builds models sequentially, with each new model attempting to correct the errors made by the previous ones. It combines multiple weak learners (typically decision trees) into a strong predictive model, excelling at capturing complex non-linear relationships. Its success in data science competitions highlights its effectiveness, though model selection requires careful consideration of data characteristics. Hist gradient boosting regressor is an optimized version of Gradient Boosting, particularly efficient for large datasets. It bins continuous features into histograms, significantly speeding up the training process. It is recognized as a powerful algorithm, for instance, in stock market prediction (Padhy et al., 2023). The final prediction aggregates the outputs of all sequentially built trees (Equation 6).

$$\hat{y}_i = \hat{y}_0 + \sum_{m=1}^{M} \eta \cdot f_m(x_i) \quad (6)$$

Extra Trees Regressor is similar to Random Forest, Extra Trees builds an ensemble of decision trees. However, it introduces more randomness by selecting split points randomly (rather than finding the optimal split) for a random subset of features at each node. This can further reduce variance and computational cost, performing well on problems with complex, non-linear relationships (Haghdoost & Md Azamathulla, 2024).

Nearest Neighbor Methods are machine learning techniques based on the similarity between examples. The most well-known method is k-Nearest Neighbors (k-NN), which classifies a new data point by examining the k closest examples in the feature space, according to a distance metric such as the Euclidean distance. The prediction is made by assigning the most common class among the nearest neighbors (for classification) or averaging their values (for regression). These techniques do not require an explicit model or intensive training, but they can be sensitive to noise and inefficient with large datasets or high-dimensional data. K-Nearest Neighbors Regressor (KNN) is a non-parametric regression algorithm based on the principle of proximity (Chacko & Chacko, 2023). To predict the value for a new instance, KNN identifies the ‘k’ closest instances (neighbors) in the training set based on a distance metric (e.g., Euclidean, Manhattan, cosine). The prediction is typically the average (or weighted average) of the target values of these ‘k’ neighbors (Equation 7).

$$\hat{y} = \frac{\sum_{i=1}^{k} w_i \cdot y_i}{\sum_{i=1}^{k} w_i} \quad (7)$$

Where w_i represents the weight assigned to neighbor i, often based on inverse distance.

Model Evaluation Metrics

Model performance was assessed using several standard regression metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Coefficient of Determination (R^2), and cross-validated RMSE (Average RMSE). These metrics quantify the difference between predicted and actual values, as defined in Equations 8-12.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{8}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{9}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{10}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2} \tag{11}$$

$$RMSE_{Average} = \frac{1}{k}\sum_{j=1}^{k}\sqrt{\frac{1}{n_j}\sum_{i=1}^{n_j}\left(y_i - \hat{y}_i\right)^2} \tag{12}$$

Here, y_i is the actual value, $\hat{y}_i$ is the predicted value, $\bar{y}_i$ is the mean of actual values, k is the number of cross-validation folds, and n is the total number of observations. R^2 indicates the proportion of variance in the dependent variable predictable from the independent variables, ranging from 0 to 1 (higher is better).

Results

This study conducts a comprehensive comparative evaluation of several regression algorithms, with a particular focus on the performance of the Gradient Boosting Regressor. The assessment utilizes key performance metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2), to provide a detailed and objective evaluation of each model's behavior on the specific dataset. These metrics offer valuable insights into the accuracy, precision, and overall fit of the models, facilitating informed decisions regarding the selection and implementation of the most appropriate machine learning techniques for the given problem. The evaluations were conducted within a cloud-based computational environment provided by Google Colab, which enabled efficient handling of large datasets and the execution of computationally intensive models.

Machine Learning Algorithms	**MAE**	**MSE**	**RMSE**	**R^2**	**Cross-Validation (Average**

					RMSE)
Multiple Linear Regression	0.65162653	0.59657800	0.77238462	0.28963023	0.78109451
Ridge Regression	0.65162561	0.59657714	0.77238406	0.28963126	0.78109446
Lasso Regression	0.74981942	0.84020996	0.91662968	-0.00047230	0.93432237
Elastic Net Regression	0.74981942	0.84020996	0.91662968	-0.00047230	0.93432237
Decision Tree Regressor	0.01534625	0.00156106	0.03951028	0.99814118	0.03477285
Random Forest Regressor	0.01235897	0.00094092	0.03067441	0.99887961	0.02810567
Gradient Boosting Regressor	0.00035279	0.00000247	0.00157250	0.99999706	0.00199047
Hist Gradient Boosting Regressor	0.01742252	0.00055091	0.02347149	0.99934401	0.02730531
Extra Trees Regressor	0.05725569	0.00579807	0.07614503	0.99309601	0.07638105
K-Nearest Neighbors Regressor	0.33556870	0.20196508	0.44940526	0.75951193	0.45980441

Table 1. Performance of Machine Learning Algorithms for Predicting Lake Titicaca Level

The results, summarized in Table 1, show the performance metrics for the ten models evaluated. The Gradient Boosting Regressor demonstrated superior performance with an MAE of 0.00035279, significantly lower than algorithms like Multiple Linear Regression (0.65162653) and K-Nearest Neighbors Regressor (0.33556870). This indicates higher average accuracy for the Gradient Boosting model. Its MSE (0.00000247) was also considerably lower than competitors like ElasticNet and Random Forest Regressor, suggesting predictions closer to the actual values with less variance in errors. The model achieved a remarkably low RMSE of 0.00157250, which is particularly informative as it represents the error in the same units as the target variable. A standout result was the R^2 value of 0.99999706, indicating the model explains nearly all the variance in the data, signifying an exceptionally good fit. Furthermore, the low average RMSE from cross-validation (0.00199047) supports the model's robustness and generalization capability.

Discussion

The findings confirm the effectiveness of the Gradient Boosting Regressor relative to other regression algorithms in this study. Its capacity to minimize MAE, MSE, and RMSE while maximizing R^2 underscores its suitability for prediction tasks demanding high accuracy.

The Gradient Boosting approach, iteratively combining weak learners into a strong ensemble, appears particularly adept at capturing complex, non-linear relationships within the dataset. This makes it advantageous in scenarios where underlying variable relationships deviate from linear assumptions common in traditional methods.

While the Gradient Boosting Regressor achieved outstanding results, data context is crucial. Models like Random Forest Regressor and Extra Trees Regressor also performed well, though not matching Gradient Boosting's metrics. However, Random Forest and Extra Trees are often less prone to overfitting and might be preferable for larger, potentially noisier datasets.

A key limitation is the potential for bias or insufficient diversity within the training dataset. Model performance in practice can vary substantially depending on the data used for training. Therefore, future research should prioritize more comprehensive validation using diverse datasets.

Finally, model interpretability is an essential consideration. Although Gradient Boosting Regressor delivered excellent technical results, understanding the basis for its predictions is often critical for real-world deployment. Techniques such as Gaussian Processes (GP), Multilayer Perceptrons (MLP), and M5P Model Trees (Wang & Wang, 2020) could be explored for enhancing model interpretability.

Conclusion

This study evaluated ten Machine Learning algorithms to predict the water level of Lake Titicaca over a thirty-year period, from 1982 to 2012. Among the models tested, the Gradient Boosting Regressor consistently outperformed the others across key evaluation metrics, such as Mean Absolute Error, Root Mean Squared Error, and R^2 score. These findings underscore the model's robustness and accuracy in capturing the complex dynamics of hydrological systems. Its strong predictive performance highlights its potential not only for supporting informed decision-making in the management of Lake Titicaca but also as a scalable approach applicable to similar water bodies worldwide. Although modeling hydrological time series presents significant challenges due to their non-linearity, seasonal variability, and data uncertainty, the application of machine learning techniques has markedly enhanced the ability to forecast water level fluctuations with greater precision. These models serve as vital tools for achieving efficient and sustainable management of water resources. Furthermore, the integration of traditional hydrological approaches with advanced Machine Learning algorithms is fostering a new paradigm in environmental modeling, offering powerful solutions for anticipating and mitigating the impacts of climate variability and human activities on aquatic ecosystems.

References

Biamont-Rojas, I. E. (2022). Heterogeneidade espacial ecotoxicológica de metais e suas implicações em reservatórios paulistas (Brasil) e para o Lago Titicaca (Peru). https://repositorio.unesp.br/server/api/core/bitstreams/4cbb2a81-0d61-4e04-8f05-297ecee531e2/content

Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32. https://doi.org/10.1023/A:1010933404324

Cabrera, J. (2012). Calibración de modelos hidrológicos. Instituto Para La Mitigación de Los Efectos Del Fenómeno El Niño, Universidad Nacional de Ingeniería, Facultad de Ingeniería Civil, Perú, 1(1).

Chacko, N., & Chacko, V. (2023). Paradigm shift presented by large language models (llm) in deep learning. Advances in Emerging Computing Technologies, 40.

Cutipa, J. M. R., Castañón, N. J. B., Larico, E. R. A., Mamani, V. S., Yucra, R. C., rro Viveros, H. P., Roque, P. Y. P., & o Macedo, W. N. (2020). Occurrences of extreme solar irradiance at 3812 meters above sea level, at Lake Titicaca (Puno-Peru). Proceedings of the LACCEI International Multi-Conference for Engineering, Education and Technology. https://doi.org/10.18687/LACCEI2020.1.1.553

Deng, B., Liu, P., Chin, R. J., Kumar, P., Jiang, C., Xiang, Y., Liu, Y., Lai, S. H., & Luo, H. (2022). Hybrid metaheuristic machine learning approach for water level prediction: A case study in Dongting Lake. Frontiers in Earth Science, 10, 928052. https://doi.org/10.3389/feart.2022.928052

Haghdoost, M., & Md Azamathulla, H. (2024). Predicting the drag coefficient of coastal trees using Support

Vector Machines and boosting ensemble models. Discover Water, 4(1), 102. https://doi.org/10.1007/s43832-024-00162-1

Hans, C. (2011). Elastic net regression modeling with the orthant normal prior. Journal of the American Statistical Association, 106(496), 1383–1393. https://doi.org/10.1198/jasa.2011.tm09241

Huang, S., Xia, J., Zeng, S., Wang, Y., & She, D. (2021). Effect of Three Gorges Dam on Poyang Lake water level at daily scale based on machine learning. Journal of Geographical Sciences, 31(11), 1598–1614. https://doi.org/10.1007/s11442-021-1913-1

Kumar, V., Kedam, N., Sharma, K. V., Mehta, D. J., & Caloiero, T. (2023). Advanced machine learning techniques to improve hydrological prediction: A comparative analysis of streamflow prediction models. Water, 15(14), 2572. https://doi.org/10.3390/w15142572

Kundzewicz, Z. W., Kanae, S., Seneviratne, S. I., Handmer, J., Nicholls, N., Peduzzi, P., Mechler, R., Bouwer, L. M., Arnell, N., & Mach, K. (2014). Flood risk and climate change: global and regional perspectives. Hydrological Sciences Journal, 59(1), 1–28. https://doi.org/10.1080/02626667.2013.857411

Lujano, E., Diaz, R. D., Tapia, B., & Lujano, A. (2023). Evaluación de Productos de Precipitación Satelital sobre la Cuenca del Lago Titicaca. Revista Brasileira de Meteorologia, 38, e38230078. https://doi.org/10.1590/0102-778638220078

Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. Journal of Applied Science and Technology Trends, 1(2), 140–147. https://doi.org/10.38094/jastt1457

Metzger Terrazas, L. (2017). Modelamiento hidrológico de la región hidrográfica del Titicaca. https://hdl.handle.net/20.500.12542/245

Mohammadi, B., Guan, Y., Aghelpour, P., Emamgholizadeh, S., Pillco Zolá, R., & Zhang, D. (2020). Simulation of Titicaca lake water level fluctuations using hybrid machine learning technique integrated with grey wolf optimizer algorithm. Water, 12(11), 3015. https://doi.org/10.3390/w12113015

Ozdemir, S., Yaqub, M., & Yildirim, S. O. (2023). A systematic literature review on lake water level prediction models. Environmental Modelling & Software, 163, 105684. https://doi.org/10.1016/j.envsoft.2023.105684

Padhy, N., Dharmireddi, S., Padhy, D. K., Saikrishna, R., & Raju, K. S. (2023). Stock Market Prediction Performance Analysis by Using Machine Learning Regressor Techniques. International Conference on Computing, Communication and Learning, 39–50. https://doi.org/10.1007/978-3-031-56998-2_4

Pathak, S., Mishra, I., & Swetapadma, A. (2018). An assessment of decision tree based classification and regression algorithms. 2018 3rd International Conference on Inventive Computation Technologies (ICICT), 92–95. https://doi.org/10.1109/ICICT43934.2018.9034296

Picado, F. (2017). Seguridad hídrica y cambio climático en la región de América Central y el Caribe: informe técnico final. Ciudad de Panamá: Centro del Agua del Trópico Húmedo para América Latina y el Caribe. https://biblioteca-repositorio.clacso.edu.ar/bitstream/CLACSO/6781/1/pdf_1480.pdf

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat, F. (2019). Deep learning and process understanding for data-driven Earth system science. Nature, 566(7743), 195–204.

Sannasi Chakravarthy, S. R., Bharanidharan, N., & Rajaguru, H. (2022). A systematic review on machine learning algorithms used for forecasting lake-water level fluctuations. Concurrency and Computation: Practice and Experience, 34(24), e7231. https://doi.org/10.1002/cpe.7231

SENAMHI. (2024). Análisis del comportamiento de la lluvia, caudales y niveles de agua en el departamento de Puno para el año hidrológico 2023 - 2024 y su pronóstico Setiembre - Noviembre. https://cdn.www.gob.pe/uploads/document/file/6886219/5951511-informe-lluvia-puno-2023-2024.pdf?v=1725413415

Shen, C., Laloy, E., Elshorbagy, A., Albert, A., Bales, J., Chang, F.-J., Ganguly, S., Hsu, K.-L., Kifer, D., & Fang, Z. (2018). HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community. Hydrology and Earth System Sciences, 22(11), 5639–5656.

Singh, V. P., Singh, R., Paul, P. K., Bisht, D. S., & Gaur, S. (2024). Machine Learning (ML) in Water Resources. In Hydrological Processes Modelling and Data Analysis: A Primer (pp. 183–202). Springer. https://doi.org/10.1007/978-981-97-1316-5_9

Sulca, J., Apaéstegui, J., & Tacza, J. (2024). New insights into the biennial-to-multidecadal variability of the water level fluctuation in Lake Titicaca in the 20th century. Frontiers in Climate, 5, 1325224.

Sulca Jota, J. C., Apaéstegui Campos, J. E., & Tacza, J. (2024). Conociendo un poco más sobre el lago Titicaca:¿ qué procesos físicos explican las variaciones del nivel de sus aguas en el tiempo?

Tan, R., Hu, Y., & Wang, Z. (2023). A multi-source data-driven model of lake water level based on variational modal decomposition and external factors with optimized bi-directional long short-term memory neural network. Environmental Modelling & Software, 167, 105766. https://doi.org/10.1016/j.envsoft.2023.105766

Wang, Q., & Wang, S. (2020). Machine Learning-Based Water Level Prediction in Lake Erie, Water, 12, 2654. https://doi.org/10.3390/w12102654

Zhu, S., Hrnjica, B., Ptak, M., Choiński, A., & Sivakumar, B. (2020). Forecasting of water level in multiple temperate lakes using machine learning models. Journal of Hydrology, 585, 124819. https://doi.org/10.1016/j.jhydrol.2020.124819

Zhu, S., Lu, H., Ptak, M., Dai, J., & Ji, Q. (2020). Lake water-level fluctuation forecasting using machine learning models: a systematic review. Environmental Science and Pollution Research, 27(36), 44807–44819. https://doi.org/10.1007/s11356-020-10917-7.