Development of a Predictive Model for Eutrophication Events using Climatic Parameters and Spatial Data in Peruvian High Andean Reservoirs

Juan A. Carlos [1], Angie Richard [2], David Ardiles [3], Felipe L. Lobo [4], Monica C. Santa-Maria[5], Ronny Gonzales [6]

[1] Fac. de Ciencias Físicas, Universidad Nacional Mayor de San Marcos, Lima, Peru - juanadriel.carlos@unmsm.edu.pe
[2] Dept. de Ingenieria Civil y Ambiental, Univ. de Ingenieria y Tecnologia (UTEC), Lima, Peru - angie.richard@utec.edu.pe
[3] TIMA, Université Grenoble Alpes – UGA, Grenoble, France – davidcesar@yahoo.com
[4] Centro de Desenvolvimento Tecnológico (CDTec), Univ. Federal de Pelotas, Pelotas, Brasil – felipe.lobo@ufpel.edu.br
[5] Dept. de Ingenieria Civil y Ambiental, Univ. de Ingenieria y Tecnologia (UTEC), Lima, Peru – msantamaria@utec.edu.pe
[6] Fac. de Ciencias e Ingenierías Físicas y Formales, Univ. Católica de Santa María, Arequipa, Peru - rgonzalesme@unsa.edu.pe

Keywords: Eutrophication, Water Quality, Predictive Modelling, NDCI, Remote Sensing, Climate Parameters.

Extended Abstract

Water quality is essential for social and economic sustainability; especially in Peru, where water scarcity and the unequal distribution of water impact various communities and economic activities. High-altitude Andean reservoirs, in particular, are a crucial water source for urban, rural, and agricultural uses, supporting ecosystems and communities throughout the Andean areas of Peru (Díaz and Sotomayor, 2013).

Eutrophication events threaten these water bodies by degrading their quality through the build-up of excess nutrients and the influence of specific climatic conditions. Climatic factors such as increased water temperature, solar radiation, wind, and rainfall can intensify eutrophication by boosting algal growth and altering nutrient cycles (Escobar and Cayetano, 2020). Ultimately, this algal overgrowth leads to reduced oxygen levels and potential toxin release, reducing water usability and posing risks to the environment and public health.

Key drivers of eutrophication in Peru include disposal of untreated wastewater, agricultural chemicals, livestock farming, inadequate disposal of solid waste, mining activities, and limited watershed management (Ojeda Solís, 2023; SINIA, 2017; Díaz Pérez, 2021; Carrasco Vela, 2020). Riparian agriculture, wastewater discharge, and nutrient-rich sediment build-up have been widely documented as contributors to eutrophication in emblematic Peruvian lakes such as Titicaca, Chinchaycocha, San Nicolás, and Conococha, leading to phytoplankton blooms, reduced water clarity, and disruptions in the food chain (SINIA, 2017; Díaz Pérez, 2021; Carrasco Vela, 2020; Díaz and Sotomayor, 2013). Furthermore, climate change and extreme weather events such as El Niño-Southern Oscillation (ENSO) can worsen hyper-sedimentation and nutrient loading to reservoirs, further threatening the sustainability of water resources in Peru (Morera et al., 2024).

This study focuses on the El Pañe reservoir, which is part of the Chili hydraulic system within the Arequipa region, where specific conditions have been identified that intensify eutrophication processes. Recent studies indicate that benthic flux — the release of nutrients from bottom sediments — exerts a significant influence on phytoplankton dynamics, accounting for 61% to 180% of chlorophyll-a variability (Mamani Larico, 2019). In addition, aquaculture, particularly fish farming, has been shown to contribute 27.7% to phytoplankton growth and 53.8% to phosphate increases in the water column (Mamani Larico, 2018). Therefore, sustainable management of the El Pañe reservoir must include sediment stabilization, regulation of aquaculture discharges, and continuous water quality monitoring, in accordance with national protocols established by the National Water Authority (ANA) (Mamani Larico, 2018; GWP, 2011). However, monitoring critical indicators such as chlorophyll and phycocyanin — key markers of algal and phytoplankton presence — poses significant challenges due to the remote and high-altitude nature of Andean water bodies. These complex geographic and climatic constraints underscore the need for more effective and accessible monitoring tools and technologies.

In this study, we developed a predictive model to identify and anticipate eutrophication events in Andean reservoirs of Peru by integrating climatic parameters with satellite spectral data. Early detection of these events will enable the timely implementation of corrective measures to mitigate their negative impact on water quality (Carrasco Vela, 2020). Our approach involved using remote sensing tools to train and refine predictive models, thus addressing the logistical and geographical challenges faced by traditional monitoring methods in the region (Redalyc, 2016). The use of satellite images, allows us to observe vast areas without the need for physical presence at the study site (Díaz and Sotomayor, 2013). This is especially valuable in remote high-altitude water bodies, where transportation is costly and fieldwork carries considerable risk. Thus, by integrating remote spectral data with climatic information, we aimed to create a robust tool for forecasting eutrophication — one that could significantly enhance water quality management across the region.

We began implementing the model by constructing a detailed data table to serve as the training set. This step was crucial, as the model's accuracy depended on the proper organization of input data for subsequent analysis. A range of key variables related to water quality and environmental conditions were collected and structured, including date, latitude, longitude, inferred chlorophyll concentration, sediments, precipitation, maximum and minimum temperatures, soil moisture, nitrogen oxide (NO_x^-) concentration, and radiation data from sources such as CFS, ERA5, and TerraClimate (Table 1). Additionally, atmospheric wind speed data from the same sources and in-situ chlorophyll concentration measurements were included to enhance the robustness of the dataset.

The model relies on Sentinel-2 imagery, from which we used the Normalized Difference Chlorophyll Index (NDCI) calculated through the AlgaeMap monitoring platform (Lobo et al., 2021). Within AlgaeMap, Sentinel-2 Level 1 images undergo atmospheric correction using SIAC (Satellite Invariant Atmospheric Correction) implemented in Google Earth Engine, which integrates MODIS MCD43A3 data and atmospheric parameters. Additional pre-processing steps include sun glint removal (Band 12 subtraction method), cloud masking with the s2cloudless product, and water body isolation using the JRC

Water Mask. After these corrections, the NDCI index is derived from Sentinel-2 Bands 5 and 4, resampled to 30 m for computational efficiency, and subsequently used in the calibration and validation of chlorophyll-a and Trophic State Index algorithms.

Table 1. Sources of climatic and satellite data used in the study.

Variable	Data Platform	Source
Satellite Images	Sentinel-2 (SR Harmonized) Sentinel-5P NOx[3]	ESA[1] – GEE[2]
Maximum Temperature	PISCO[4]	SENAMHI[5]
Minimum Temperature	PISCO	SENAMHI
Sunshine Hours	PISCO	SENAMHI
Precipitation	PISCO	SENAMHI
Solar Radiation	CFSV2: NCEP[6] ERA5 TerraClimate	NOAA[7] ECMWF[8] USGS[9] and Idaho EPSCoR
Wind Speed	CFSV2: NCEP ERA5 TerraClimate	NOAA ECMWF USGS and Idaho EPSCoR

[1] European Space Agency.
[2] Google Earth Engine.
[3] Based on nitrogen dioxide levels.
[4] Interpolation System of Climatological and Hydrological Data.
[5] National Service of Meteorology and Hydrology of Peru.
[6] Climate Forecast System.
[7] U.S. National Oceanic and Atmospheric Administration.
[8] European Centre for Medium-Range Weather Forecasts.
[9] U.S. Geological Survey.

The data table was built using pixel values extracted from raster files of the Normalized Difference Chlorophyll Index (NDCI) for the El Pañe reservoir, with a spatial resolution of 120 meters, obtained from the AlgaeMap platform. Each pixel's data was organized by location and date, assigning corresponding values for the previously mentioned variables. This structure enabled each row in the table to represent a unique observation with all the necessary attributes for processing by machine learning models. Inferred chlorophyll-a concentrations were derived from the NDCI using a nonlinear empirical model based on a power-law fit developed by Lobo et al. (2021), expressed by the following equation:

$$Chl - a = 23.44 \times (NDCI + 1)^{7.95} \qquad (1)$$

The accuracy of the model for estimating chlorophyll-a concentrations at the El Pañe Reservoir was assessed using the root mean square error (RMSE), by comparing model-derived values with *in situ* measurements obtained from previous studies (Figure 1).

Pixels from the spectral data were classified into two groups: those within the reservoir's water body and those in the surrounding terrestrial areas. To estimate sediment contributions, pixels within the water body were assigned null or zero values, as they are not directly influenced by sedimentation processes. In contrast, pixels in the peripheral zones were affected by sedimentation and therefore exhibited varying soil loss values. This classification approach enabled more accurate modelling by accounting for the interactions between aquatic and adjacent terrestrial environments. To estimate sediment levels in peripheral areas, we applied the RUSLE (Revised Universal Soil Loss Equation), adapted to each zone, delimited according to the terrain's topography. By applying the equation to all pixels within each defined area, we were able to calculate average sediment loss as a function of various topographic and climatic variables.

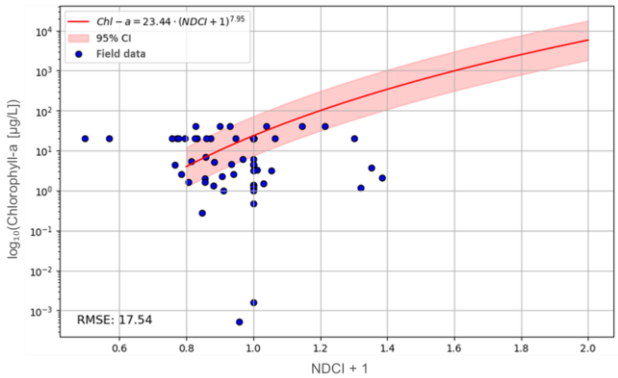

Figure 1. Correlation between chlorophyll concentrations estimated using equation 1 (red line) and historic field data from el Pañe Reservoir.

In constructing the data table, the average chlorophyll value of adjacent neighbouring pixels in the cardinal directions (north, south, east, and west) was included for each target pixel. This average was calculated using only valid values, excluding null (NA) or zero entries, to prevent distortions in the analysis. This approach was designed to mitigate the "salt and pepper" effect, which can occur when chlorophyll concentration predictions display abrupt or unrealistic variations. By incorporating the average values of adjacent pixels, we aimed to smooth the predictions and promote gradual transitions between observed and predicted values, thereby enhancing the overall accuracy and reliability of the model outputs.

After constructing the data table, a Random Forest model was trained with the following parameters: 200 trees, a maximum depth of 15, a minimum of two samples for splitting nodes, and a minimum of 10 samples per leaf. The response variable was the inferred chlorophyll-a concentration (from equation 1). However, the model struggled to adequately predict high chlorophyll-a values, as shown in Figure 2, where notable discrepancies appear between the "observed" values (blue lines) and the model-predicted values (red lines).

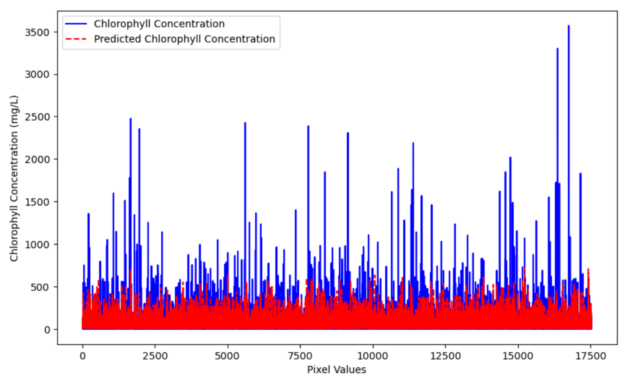

Figure 2. Correspondence between chlorophyll concentration values estimated from NDCI data using equation 1 (blue) and estimated using the Random Forest model in this study (red).

This outcome led us to shift towards a classification model that groups chlorophyll-a concentrations based on the Trophic State Index (TSI). The new model categorized chlorophyll-a values into five groups: oligotrophic (< 3.24 µg/L), mesotrophic (< 11.03 µg/L), eutrophic (< 30.55 µg/L), super-eutrophic (< 69.05 µg/L), and hyper-eutrophic (> 69.05 µg/L). Additionally, a binary classification was implemented to detect the presence of algal blooms by identifying water bodies as eutrophic or non-eutrophic, depending on whether chlorophyll-a concentrations exceeded 30.55 µg/L. This approach significantly enhanced the model's performance in predicting and classifying eutrophication conditions and algal blooms in the El Pañe reservoir.

Model validation was conducted using several performance metrics, including accuracy, a confusion matrix, a classification report, and Receiver Operating Characteristic (ROC) curve analysis. The first ROC curve corresponds to the binary classification model used to detect the presence or absence of algal blooms (Figure 3). This curve yielded an Area Under the Curve (AUC) of 0.92, indicating excellent discrimination between the "Yes" and "No" classes (i.e., presence or absence of algal blooms) (Hosmer & Lemeshow, 2000).

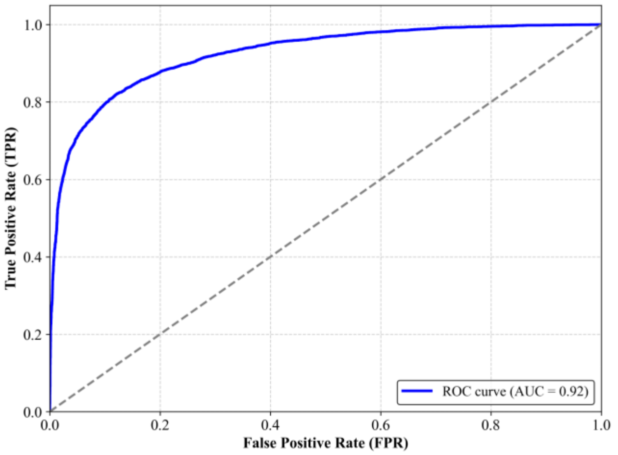

Figure 3. ROC curve of binary classification of eutrophication conditions (eutrophic or non-eutrophic) in el Pañe reservoir, as predicted by our TSI model.

As shown in the curve, the model achieved a True Positive Rate (TPR) close to 1 with a very low False Positive Rate (FPR), demonstrating a strong ability to accurately identify eutrophic water bodies. In terms of precision and recall metrics, the model reached an accuracy of 0.86, further supporting the reliability of its predictions in the binary classification scenario (Fawcett, 2006).

In addition, the multi-class classification based on Carlson's Trophic State Index (TSI) was evaluated using ROC curves for each TSI category. The results, shown in Figure 4, illustrate the model's performance across five trophic levels: oligotrophic, mesotrophic, eutrophic, super-eutrophic, and hyper-eutrophic. The ROC curves yielded AUC values ranging from 0.72 to 0.82, with the oligotrophic class achieving the highest performance (AUC = 0.82). The hyper-eutrophic and super-eutrophic classes also showed strong performance, with AUC values of 0.81 and 0.72, respectively. In contrast, the mesotrophic and eutrophic categories exhibited slightly lower AUCs, indicating the model's reduced ability to distinguish between these intermediate TSI ranges (Figure 4). Despite these differences, the model was able to correctly classify chlorophyll concentrations across the various TSI categories, demonstrating satisfactory performance. This outcome underscores the value of TSI as an effective tool for assessing water quality (Carlson, 1977).

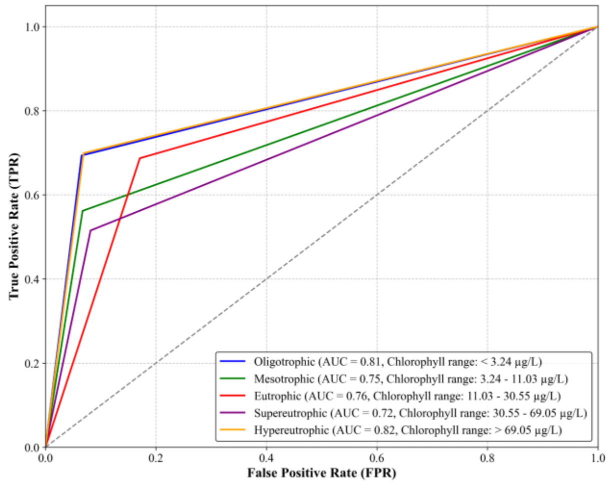

Figure 4. ROC curves showing the model performance across five TSI categories: oligotrophic, mesotrophic, eutrophic, super-eutrophic, and hyper-eutrophic.

To assess the relative importance of the variables used in the Random Forest models, a Principal Component Analysis (PCA) was conducted. As shown in the cumulative variance plot (Figure 5), the first 11 principal components explain nearly all of the variance in the dataset. This indicates that these components retain most of the original information, and highlight PCA's effectiveness in reducing dimensionality. Furthermore, by concentrating essential variability into fewer components, PCA enhances the model's computational efficiency and interpretability while minimizing multicollinearity among variables (De Marco et al., 2018).

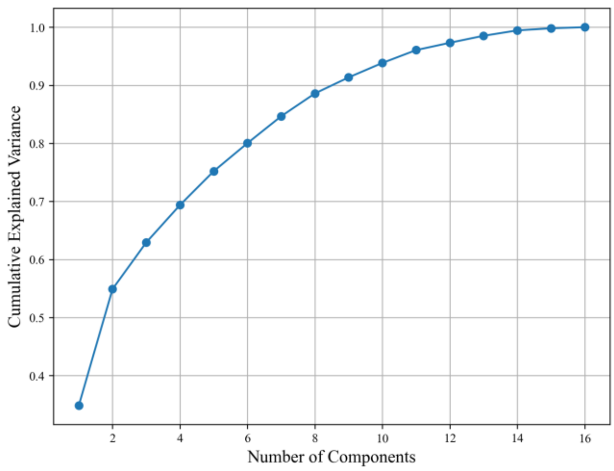

Figure 5. Principal Components Analysis (PCA) of the variables used in the random forest models.

Regarding the contributions of individual variables to each principal component, the analysis revealed that certain components are more strongly influenced by specific variables (Figure 6). For instance, variables related to radiation (Rad_CFS and Rad_TC) and wind speed (Wind_CFS_u, Wind_CFS_v) were dominant in the first principal components, particularly PC1 and PC2. This suggests that climatic and meteorological factors may play an important role in influencing chlorophyll concentration. In contrast, components such as PC6 and PC7 were more influenced by variables like soil moisture (HS) and maximum temperature. In any case, this detailed breakdown of variable contributions across principal components helps identify the key drivers of the trophic state in the El Pañe reservoir and may contribute to a more effective management of this water resource.

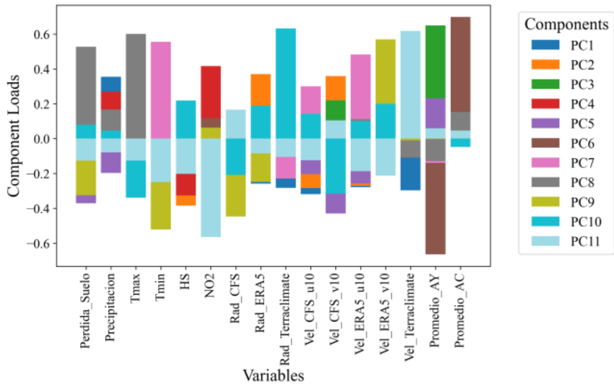

Figure 6. Contribution of individual variables (in the Random Forest model) to principal components in the PCA analysis.

Overall, the predictive models developed in this study offer valuable applications for water quality management, particularly in the classification of water bodies according to the Carlson Trophic State Index (TSI). This index enables the categorization of water bodies into trophic levels, which is essential for detecting and managing eutrophication and algal blooms (Carlson, 1977). By accurately classifying chlorophyll-a concentrations, the model helps identify ecosystems in need of intervention to address biodiversity loss and declining water quality. Additionally, the binary algal bloom detection model enhances the ability to assess risks to potable water sources — an important consideration for the drinking water industry (Paerl & Huisman, 2009). In the environmental management sector, the model supports public policy planning by enabling authorities to prioritize restoration and protection efforts for vulnerable water bodies (Smith et al., 1999). It also has practical applications in agriculture and aquaculture, helping to monitor and maintain water quality for irrigation and cultivation, preventing issues related to pollution and algal overgrowth (Lürling et al., 2013). Moreover, in drinking water treatment, the model contributes to greater operational efficiency by predicting water quality and informing process adjustments to ensure compliance with safety standards (Schindler, 2006). By enabling the identification of high-risk areas and forecasting future eutrophication events, this model could provide a proactive tool for effective water resource management, and to significantly enhance water quality management in Peru's Andean reservoirs through data-driven decision-making and timely interventions.

In this study, we provided a comprehensive assessment of the trophic status of the El Pañe reservoir using machine learning techniques — specifically, a Random Forest classification model — and Principal Component Analysis (PCA) to identify key variables influencing chlorophyll concentration. Our findings show that the classification approach was more effective than the regression approach, which struggled to predict high chlorophyll-a values, showing discrepancies between "observed" and predicted values (Figure 2). It is important to emphasize that the accuracy of the Lobo et al. (2021) model for estimating "observed" chlorophyll concentrations needs further validation in Andean reservoirs. In contrast to the Random Forest models, the classification model offered a more accurate representation of eutrophication conditions, achieving an AUC of 0.92 in the binary classification of algal bloom presence, and acceptable AUC values across the various chlorophyll concentration categories.

Principal Component Analysis (PCA) was used to reduce model complexity and enhance the interpretability of the variables. The first 11 principal components accounted for a significant portion of the variance in the dataset, enabling the identification of key influencing factors such as radiation, wind speed, and soil moisture. The variable loadings on these components revealed that climatic and meteorological factors — particularly radiation and wind speed — play a critical role in influencing chlorophyll concentrations in the reservoir, offering valuable insight into the dynamics of algal blooms and overall water quality.

Based on the results presented, it can be concluded that the developed model, based on TSI, is effective in predicting the trophic state of the El Pañe reservoir, enhancing water resource management through accurate classification of eutrophication conditions. Model validation demonstrated strong performance, with an accuracy of 85.9%, outperforming a baseline model. Additionally, the application of PCA helped identify the most influential variables driving chlorophyll concentrations, offering valuable insights into the ecological dynamics of the reservoir. Overall, our research contributes to the prediction of water quality, and presents a robust methodology for analysing large-scale environmental datasets.

References

Carlson, R. E., 1977: A trophic state index for lakes. Limnology and Oceanography, 22(2), 361-369.

Carrasco Vela, D. L., 2020: Nivel de eutrofización de la laguna San Nicolás (Cajamarca, Perú) mediante teledetección. Colegio de Geógrafos del Perú.

De Marco P Júnior, Nóbrega, C. C., 2018: Evaluating collinearity effects on species distribution models: An approach based on virtual species simulation. PLoS ONE, 13(9), e0202403. doi.org/10.1371/journal.pone.0202403.

Díaz Pérez, C., 2021: Efectos de la actividad agrícola ribereña en el proceso de eutrofización de la laguna Pucush Uclo en Chupaca. Repositorio Universidad Continental.

Díaz y Sotomayor, 2013: Evaluación de la eutrofización de la laguna Conococha. Universidad de Viena.

Fawcett, T., 2006: An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874. doi.org/10.1016/S0269-7491(99)00091-3.

GWP, 2011: Protocolo de Monitoreo de la Calidad de los Cuerpos de Agua. Global Water Partnership.

Hosmer, D. W., & Lemeshow, S., 2000: Applied Logistic Regression (2nd ed.). Wiley, New York.

Lobo, F. D. L., Nagel, G. W., Maciel, D. A., Carvalho, L. A. S. D., Martins, V. S., Barbosa, C. C. F., & Novo, E. M. L. D. M., 2021: AlgaeMAp: Algae bloom monitoring application for inland waters in Latin America. Remote Sensing, 13(15), 2874. doi.org/10.3390/rs13152874.

Lürling, M., Van Oosterhout, F., & Faassen, E., 2013: The role of nitrogen and phosphorus in promoting harmful cyanobacterial blooms. Aquatic Ecology, 47(3), 295-306. doi.org/10.1016/j.hal.2016.01.010.

Mamani Larico, A. J., & Zúñiga Medina, S. A., 2019: Application of WASP model for assessment of water quality for eutrophication control for a reservoir in the Peruvian Andes. Lakes & Reservoirs: Research & Management, 24(1), 37-47. doi.org/10.1111/lre.12256.

Mamani Larico, A. J., 2018: Implementación del modelo numérico de calidad del agua WASP para la evaluación del fenómeno de eutrofización del embalse El Pañe. Universidad Nacional de San Agustín de Arequipa.

Morera, S., Foucher, A., Sanchez, M., Orrillo, J., & Evrard, O., 2024: Cuando el clima y las interacciones humanas amenazan los recursos suelo y agua a través de la hipersedimentación: caso del embalse de Poechos, Perú. Instituto Geofísico del Perú. doi.org/10.5194/hess-27-3191-2023.

Ojeda Solis, D., 2023: Evaluación de las condiciones hídricas y su relación con la calidad de agua en el distrito de Carhuamayo, Junín-2021. Repositorio Universidad Continental. https://hdl.handle.net/20.500.12394/13734.

Paerl, H. W., & Huisman, J., 2009: Climate change: A catalyst for global expansion of harmful cyanobacterial blooms. Environmental Microbiology Reports, 1(1), 27-37. doi.org/10.1111/j.1758-2229.2008.00004.x.

Schindler, D. W., 2006: Recent advances in the understanding and management of eutrophication. Limnology and Oceanography, 51(1), 14-18. doi.org/10.4319/lo.2006.51.1_part_2.0356.

SINIA, 2017: Estado del Ambiente y los Impactos. Sistema Nacional de Información Ambiental.

Smith, V. H., Tilman, G. D., & Nekola, J. C., 1999: Eutrophication: Impacts of excess nutrient inputs on freshwater, marine, and terrestrial ecosystems. Environmental Pollution, 100(1-3), 179-196. doi.org/10.1016/S0269-7491(99)00091-3.