Constructing Homogeneous Water Quality Time Series Through Multivariate Modeling: A Case Study of the Paraíba do Sul River Basin

Abstract: This study aims to create a homogeneous time series of water quality parameters in the Paraíba do Sul River Basin – Rio de Janeiro, Brazil. The basin has high socioeconomic and environmental relevance for the region, being intensively used for urban water supply and industrial and agricultural activities. The analyzed parameters were pH, dissolved oxygen, and turbidity. The time series of these variables present substantial temporal heterogeneity, varying according to the parameter and the monitoring station analyzed. Five gauging stations in the Paraíba do Sul River Basin were evaluated, using data obtained from Hidroweb of the National Water Agency (ANA), covering the period from 2000 to 2024.

The homogenization was performed on a monthly basis through the application of XGBoost models, using as predictor variables the last ten measurements of each water quality parameter, the observed precipitation, and the ZCAS and atmospheric blocking indices during the three months preceding the month to be estimated. Model performance was assessed using cross-validation techniques, in which the time series was divided into five equally sized blocks, with one block used for model validation at each iteration. The error metrics applied were RMSE and MAE, in order to evaluate the average error of the model estimates for all analyzed variables.

This methodological approach constitutes a powerful tool for constructing homogeneous water quality time series, as it is based on a multivariate framework, although it does not consider other non-environmental factors, such as anthropogenic influences, which could also affect these parameters.

keywords:Cross-validation, ERA5 reanalysis, Atmospheric blocking, South Atlantic Convergence Zone (SACZ), Predictive modeling