Goldschmidt 2026 Abstract
https://doi.org/10.46427/gold2026.33887

Development of a Data-Driven Water Quality Index for the Yamuna River Using Random Forest Modeling Approach

ABDUL GANI[1], SHRAY PATHAK[2] AND ATHAR HUSSAIN[1]

[1]Netaji Subhas University of Technology
[2]Indian Institute of Technology Ropar

The Yamuna River is an important part of northern India's socioeconomic and cultural fabric. Unfortunately, its water quality has significantly declined due to industrialization, agriculture, and rapid urbanization. Accurate assessment of surface water quality is critical for effective management of freshwater resources, particularly in densely populated and industrialized river basins such as the Yamuna River in northern India. By utilizing the Water Quality Index (WQI), complex data on water quality was reduced to a single number, which made it easier to understand the river's general health and suitability for various applications. The present study proposes a data-driven approach for the development and prediction of the Yamuna River's WQI by using the Random Forest (RF) machine learning model. The data of various physicochemical parameters of different monitoring stations on the Yamuna River was collected from the Central Water Commission for a period of 15 years. The WQI was first calculated by using a hybrid aggregation technique and used as the targeted variable for the model training. The obtained WQI varied from 4.96 to 88.48, in which the minimum value was obtained at Agra and the maximum value was obtained at Mawi. The RF model was used to describe nonlinear interactions between the corresponding WQI values and water quality factors. The model's performance was assessed by mean absolute error (5.06), root mean square error (8.71), and coefficient of determination (0.74), demonstrating high prediction accuracy. The developed RF-based WQI model provides a strong framework for decision assistance in river basin management and real-time water quality evaluation. The suggested model provides a strong decision-support tool for environmental risk assessment, policymaking, and real-time water quality forecasting, especially when applied to dynamic and anthropogenically stressed riverine systems.