26th International Congress on Modelling and Simulation, Adelaide, SA, Australia, 30 Nov–4 Dec 2025
mssanz.org.au/modsim2025

A hybrid decomposition and machine learning model for forecasting coastal chlorophyll-a and total nitrogen

Xiaotong Zhu [a], **Minghua Zhou**[a]

[a] *Key Laboratory of Mountain Surface Processes and Ecological Regulation, Institute of Mountain Hazards and Environment, Chinese Academy of Sciences, Chengdu, 610213, China*
Email: zhuxiaotong@imde.ac.cn

Abstract: Chlorophyll-a (Chla) and Total Nitrogen (TN) concentrations are key indicators for the early detection of algal blooms, essential for managing coastal ecosystems (Mohebzadeh et al., 2020). However, the optical complexity of coastal waters complicates the accurate forecasting of these parameters. To address this, we propose a novel hybrid model, CEEMDAN-VMD-XGBoost-SARIMA (CVXS), designed to improve forecasting accuracy for Chla and TN concentrations.

The CVXS model combines several advanced methods to improve forecasting performance. First, the Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) method is applied to decompose the original time series data of Chla and TN into a set of intrinsic mode functions (IMFs). This decomposition is particularly effective for isolating different frequency components and removing noise, which is crucial for modelling complex and non-linear time series data. To further refine the decomposition, the Variational Mode Decomposition (VMD) technique is used to break down the high-frequency IMFs (IMF_1 and IMF_2) into multiple modes, each representing a distinct oscillation pattern. This approach allows for better separation of multi-scale features in the data, particularly those related to rapid fluctuations in water quality indicators.

For the prediction step, we utilize a combination of machine learning and time series models. The Extreme Gradient Boosting (XGBoost) algorithm is employed to model the intermediate-frequency IMFs and modes, taking advantage of its capability to handle complex non-linear relationships and reduce overfitting. For the low-frequency IMFs, the Seasonal Autoregressive Integrated Moving Average (SARIMA) model is applied, which is particularly effective in capturing seasonal trends and long-term dependencies in time series data. By integrating XGBoost and SARIMA for different frequency components, the CVXS model offers a robust and balanced approach to forecasting.

The CVXS model was validated using data from three hydrological monitoring stations in Hong Kong, China, and compared with benchmark models, including XGBoost, support vector regression (SVR), Multilayer Perceptron (MLP), SARIMA, and three hybrid models: CEEMDAN-XGBoost, CEEMDAN-SVR, and CEEMDAN-MLP. Results indicated that the CVXS model consistently outperformed all other models in forecasting both ChlA and TN concentrations, with NSE values remaining above 0.97 across all stations, while CEEMDAN-XGBoost ranked second with NSE ranging from 0.80 to 0.90, and the single models (XGBoost, SVR, MLP, SARIMA) generally showed lower accuracy with NSE often below 0.70. In particular, the CVXS model achieved the highest accuracy for short-term forecasts, with an optimal forecasting horizon of one to two months. The model's ability to capture the temporal dynamics of ChlA and TN is further supported by a correlation analysis between the decomposed time series and hydrological factors such as precipitation, temperature, and river discharge. The results suggest that these hydrological factors are inherently embedded in the decomposed data, allowing the model to accurately predict fluctuations in ChlA and TN concentrations that are influenced by environmental changes.

Overall, the findings highlight the promising potential of the CVXS model in improving the accuracy of algal bloom forecasting and enhancing coastal water resource management. This model can serve as an effective tool for early-warning systems, contributing to the sustainable management of coastal ecosystems facing eutrophication and other water quality challenges.

REFERENCES

Mohebzadeh, H., Yeom, J., Lee, T., 2020. Spatial Downscaling of MODIS Chlorophyll-a with Genetic Programming in South Korea. Remote Sensing 12.DOI:10.3390/rs12091412

Keywords: *Water quality, machine learning, decomposition algorithm, extreme gradient boosting algorithm, forecasting*

