Using a boundary-corrected wavelet transform coupled with machine learning and hybrid deep learning approaches for multi-step water level forecasting in Lakes Michigan and Ontario

Rahim Barzegar\textsuperscript{1,2}, Jan Adamowski\textsuperscript{1}, John Quilty\textsuperscript{3}, and Mohammad Taghi Aalami\textsuperscript{2}

\textsuperscript{1}Department of Bioresource Engineering, McGill University, Montreal, Canada (rahim.barzegar@mail.mcgill.ca)
\textsuperscript{2}Department of Water Resources Management, University of Tabriz, Tabriz, Iran
\textsuperscript{3}Department of Civil and Environmental Engineering, University of Waterloo, Waterloo, Canada

Accurate water level (WL) forecasting is important for water resources management and planning purposes in the Great Lakes. The objectives of this research are two-fold. The first objective is to apply machine learning (ML) (i.e., random forest (RF) and support vector regression (SVR)) and hybrid convolutional neural network (CNN)-long short term memory (LSTM) deep learning (DL) models for multi-step (i.e., one-, two- and three-monthly step ahead) WL forecasting in the Great Lakes (Michigan and Ontario). The second objective is to integrate the boundary corrected (BC) maximal overlap discrete wavelet transform (MODWT) with SVR, RF, and CNN-LSTM models to improve the performance of the individual models. By employing a BC-wavelet decomposition method, the ‘future data’ issue (i.e., data from the future that is not available), often overlooked in the literature and a major barrier to achieving realistic forecasting performance is overcome.

For Lakes Michigan and Ontario, 1212 monthly WL (m) records (spanning Jan 1918–Dec 2018) were used to develop the models. For the non-wavelet-based models (SVR, RF, and CNN-LSTM), candidate model inputs included the WL recorded over the previous 12 months. For the BC-MODWT-based models (BC-MODWT-SVR, BC-MODWT-RF, and BC-MODWT-CNN-LSTM), the lagged input time series were decomposed into BC-wavelet and scaling coefficients by using different mother wavelets (Haar, Daubechies, Symlets, Fejer-Korovkin and Coiflets), filter lengths (from two up to 12) and decomposition levels (from one up to seven). For each method (SVR, RF, and CNN-LSTM), mother wavelet, and decomposition level a model was generated. For both wavelet- and non-wavelet-based models, the particle swarm optimization (PSO) method was used to select the most appropriate inputs to include in the proposed multi-step WL forecasting models.

The datasets were partitioned into calibration and validation subsets. After calibrating the models, various performance evaluation metrics, e.g., coefficient of determination ($R^2$), root mean square error (RMSE), mean absolute error (MAE), root mean square percentage error (RMSPE), mean absolute percentage error (MAPE) and the Nash-Sutcliffe efficiency coefficient (NSC) were used to assess model accuracy.

Of the ML models, the SVR outperformed RF while the DL models outperformed the ML models.
for each forecast lead time (one-, two-, and three-step(s) ahead). Results from this case study indicate that not all wavelet families and decomposition levels perform equally and in some cases, the wavelet-based models do not improve performance over the non-wavelet-based models. However, the BC-MODWT-CNN-LSTM using suitable mother wavelets (e.g., Haar) outperforms the individual ML and BC-MODWT-ML-based models. More accurate forecasts were obtained for Lake Michigan although the performance in both Great Lakes was accurate. The outcomes of this research indicate that the BC-MODWT-CNN-LSTM model is a promising tool for generating accurate WL forecasts.