Abstract

Anthropogenic activities have resulted in substantial growth of lake eutrophication cases all over the globe from the last century. It is a well-known fact that eutrophication is a process of unreasonable nutrient advancement, by and large with phosphorus and nitrogen to the waterbody, which brings about a significant deterioration in the aquatic ecosystem. Eutrophication is related to the production of toxic algal blooms, decreased oxygen concentrations, and increased turbidity in the waterbody. These phenomena can prompt serious wellbeing peril as far as drinking water quality issues are concerned and have unfavourable social and economic effects. Considering such adverse effects, management and control of eutrophication gets unavoidable.

From the last four decades several modelling methodologies have been developed to forecast and mitigate lake eutrophication with various levels of accomplishment. Application of both process-based and data-driven modelling approaches have been observed frequently in lake modelling. However, the process-based models rely on extensive input data, expertise, and computational resources, and often based on simplifications and assumptions which may affect their accuracy. These factors lead to popularity of data-driven approaches in lake modelling. Data-driven modelling approaches, like the artificial neural network (ANN), are being increasingly utilized in ecological modelling in recent times. Such machine learning based supervised models can regenerate the non-linear relationship between the ecosystem parameters in a more efficient manner, unlike the multiple regression techniques, which generally consider linear relationship among variables for model development. Moreover, as the underlying relationships among ecological variables are commonly intricate and accompany extensive uncertainties, machine learning based data-driven models are developing as a useful modelling tool for ecosystem management and restoration. Apart from neural networks, use of other data-driven modelling techniques such as support vector regression (SVR), Adaptive neuro-fuzzy inference system (ANFIS), Gaussian process regression (GPR) etc. are gaining attention in lake modelling. Water quality parameters like dissolved oxygen (DO), total nitrogen (TN), total phosphorus (TP), chlorophyll-a (Chl-a), Secchi depth (SD) etc. are commonly used as eutrophication indicators in these models.

Nevertheless, collection of an exhaustive dataset for successful training of a predictive data-driven model is a major concern. It is seen that lake eutrophication models using data-driven approach were developed with water quality parameter datasets collected over a long duration of time like ten to twenty years. So, under the circumstances where such prolonged data is unavailable for water bodies, which is most likely in remote areas in underdeveloped and developing countries, eutrophication management becomes irrational with data-driven modelling. Assam, the North-Eastern state of India is bestowed with large number of small to large water bodies. Urbanisation, improper dumping of sewage and other anthropogenic activities have made the surface waterbodies vulnerable to eutrophication in this region. The number of research work highlighting the water quality in lakes of Assam and its trophic state are very limited and periodic water quality monitoring data is also not available. Considering these factors, the major objective of the present work was to study the comprehensive lake eutrophication dynamics by simulating artificial lake eutrophication scenario and its subsequent effective restoration mechanism with the help of predictive data-driven modelling.

Three model tanks were constructed, two concrete tanks (Tank 1 and Tank 2) and one artificial pond (Tank 3) to simulate lake eutrophication scenario. These prototype lakes were initially filled with clear water and thereafter periodic application of wastewater was done until the hypereutrophic state was achieved. In case of Tank 1 and Tank 2 eutrophication process was allowed to occur in a controlled environment whereas Tank 3 was replicated like a natural pond. Tests for water quality parameters were regularly conducted on the studied prototype lakes. Changes in the water quality of the investigated lakes were verified with trophic status index (TSI). The investigated dataset were analysed statistically and was used for development of data-driven models for prediction of eutrophication indicators in the water bodies in Assam, India. The trained models were subsequently tested against natural water body data from samples collected from six different locations in Assam. To study the effect of each independent variable on the prediction of dependent variables, sensitivity analysis was done based on data perturbation technique.

The process of experimental investigation on the prototype lakes was done in two trials. In the first trial, Tank 1 and Tank 2 were used as prototype lakes and pH, TDS, EC, BOD, DO, turbidity, TN, TP, SD, and water temperature were monitored periodically. Based on this dataset ANN, multiple linear regression, GPR, and SVR methods were used to model

eutrophication indicators DO and SD. Under ANN approach, two methods namely multi-layer perceptron (MLP) and time delay neural network (TDNN) were used to model the target parameters. The inputs of the models were determined based on a model-based data-pruning method in MLP architecture and thereafter the same inputs were used for other models. The optimum architecture of the ANN, SVR, and GPR models were determined by adjusting the hyperparameters of the models on a trial and error method. Overall, five types of DO and SD models were trained and tested based on the considered machine learning (ML) algorithms. The efficiency of model prediction was evaluated based on statistical indices coefficient of determination (R^2), Nash-Sutcliffe efficiency (E), root mean square error (RMSE), and mean absolute error (MAE). In the second trial Tank 1 and Tank 3 were used to simulate the process of eutrophication and Chl-a was also periodically monitored in addition to the other water quality parameters. Dataset gathered from each prototype lake was used to develop two DO, SD, and Chl-a models respectively using ANN, GPR, and ANFIS approach. Total six numbers of DO, SD, and Chl-a models were trained and tested respectively based on the two prototype lake's data and three ML algorithms considered. The input selection, model architecture selection, model efficiency evaluation criteria were similar to the first trial.

Lake eutrophication scenario was successfully replicated by continuous addition of nutrient rich domestic waste water into artificially constructed prototype lakes during both the trials. Trophic status index (TSI) was estimated for the studied lakes and it was found that average TSI values were greater than 70 during the final stages of investigation. It indicated transition of the lakes from a clear water oligotrophic to hypereutrophic stage. pH, EC, TDS, turbidity, TN, TP, Chl-a, and BOD of the lakes were found to increase considerably during the period of investigation and decrease in DO, SD values were observed. The linear correlation among investigated water quality parameters were found to be poor as reflected by smaller correlation coefficient values of less than 0.60 between parameters. From the model training and testing results of the first trial it was seen that prediction accuracy of linear regression model was very poor compared to other methods. MLP, TDNN and GPR based DO and SD models showed better goodness-of-fit between model predicted and actual values with less errors compared SVR models. R^2 and E values greater than 0.95 was observed for the DO and SD prediction in case of MLP, TDNN and GPR model training. During model testing against natural waterbody, these models showed acceptable prediction accuracy and R^2 value greater than 0.90 was achieved. As ANN and GPR models were found

to have better prediction accuracy, in the second trial also these two methods along with sophisticated ANFIS algorithm was considered for eutrophication modelling. For DO, SD, and Chl-a prediction 5, 5, and 9 numbers of input parameters were optimized. Results of the second trial revealed that ANN, GPR, and ANFIS models were able to predict all the target variables with better accuracy during both model training and testing phase where R^2 of 0.99 and greater than 0.94 was observed respectively. R^2, E, RMSE, and MAE values were found to be superior for the optimum models under second trial compared to first trial. Model testing with natural water body data in both trials revealed suitability of adopted experimental and modelling approach for lake eutrophication prediction. Major sensitive parameters for DO prediction was found as temperature and Chl-a. pH, increase in nutrient concentration, DO, and increase in temperature were reported as most sensitive for Chl-a estimation. The sensitivity of inputs were less for SD model.

Out of different data-driven modelling approaches investigated ANN, GPR, and ANFIS were found to predict the eutrophication indicators with better accuracy. Overall, ANFIS models were found to be more accurate and robust. R^2 value of 0.99 has been achieved for ANFIS models during both model training and testing phase. Out of the two types of prototype lakes considered, models based on artificial pond data were found to produce slightly better prediction accuracy. But data gathered from concrete tanks were less time consuming and modelling results were also promising. The presented models were successful in predicting eutrophication indicators in natural water bodies in Assam, India. This type of data-driven modelling approach based on laboratory investigated data on artificially simulated lake systems could be an alternate solution to rapid eutrophication management of water bodies where prolonged water quality data is unavailable for the same.

Keywords: ANN, ANFIS, Dissolved oxygen, Data-driven models, Lake eutrophication, Secchi Depth, Chlorophyll-a, Gaussian process regression, Support vector regression.