CHAPTER 2
LITERATURE REVIEW

This chapter provides a comprehensive overview of the available literature in the field of lake eutrophication modelling. Initially the present condition of lake eutrophication pertaining to global scenario has been presented. The background of lake modelling and advances in mathematical process-based models have been discussed in the next sections. The advances in lake management through data-driven modelling approaches with special emphasis on the considered model types of viz. ANN, ANFIS, SVM, and GPR under this study has been discussed. Finally major conclusions of the literature review and research gap has been addressed.

2.1 GLOBAL SCENARIO OF LAKE EUTROPHICATION

Lake eutrophication is a widespread and persistent problem that affects freshwater systems around the world. Lake eutrophication is a global phenomenon, with many lakes across the world affected by excessive nutrient loading. Some of the most eutrophic lakes in the world are found in densely populated areas of Europe, North America, and Asia, where the agricultural and industrial activities are intense. According to a study conducted by the World Resources Institute, approximately 54% of the world's lakes are experiencing eutrophication, with the highest levels of eutrophication found in Asia, Europe, and North America [40]. The study also found that agricultural runoff, sewage discharge, and urbanization were the leading causes of eutrophication in lakes. According to the survey of International Lake Environment committee (ILEC) [72], eutrophic lakes make up 54% of lakes in the Asia Pacific region; the corresponding percentages in Europe, Africa, North America, and South America are 53%, 28%, 48%, and 41%. Tremendous increase in human population in the last century have converted eutrophication from local issue to a global concern [189]. So, due to degradation of water quality as a result of eutrophication in lakes, rivers, and streams, limnological studies have been emphasized in recent times [159].

Lake Erie, one of the North America's Great Lakes is a prime example of eutrophication caused by manmade activities. The lake (388 kms long and 92 kms wide) has a maximum and average depth of 64 and 19 meters respectively. Since the lake is shallow and warm, and it is highly rich in nutrients, the lake is biologically the most productive [147]. Nutrient enrichment mainly with phosphorus in the lake was the major cause of massive blue-green algae blooms. Because of these blooms, less light penetrated to the deeper parts of the lake, reducing rate of photosynthesis and oxygen production by other phytoplankton groups. These blooms have resulted in beach closures and a significant impact on the local tourism industry, affecting the economic wellbeing of the region [160].

In Asia, eutrophication is a growing problem, particularly in China and India, where rapid economic growth and population growth have led to increased nutrient inputs into water bodies. Lake Taihu, located in China's Jiangsu province, has experienced severe eutrophication due to industrialization and population growth in the surrounding area. The lake's nutrient levels are four times higher than the limit set by the World Health Organization. This has led to harmful algal blooms and water quality issues, impacting the health and livelihoods of the local population [79, 192]. Eutrophication cases has been reported in 67 lakes in China, accounting for approximately 51.2% of the total number of lakes. According to estimates of Yang et al. [199], about 30 billion tons of polluted water will be discharged directly into the lakes by 2030. Therefore, eutrophication may be a concern for all of China's urban lakes as well as the majority of the medium sized lakes in the country's urban-rural boundary zones.

In India, majority of lakes and urban water bodies are also experiencing the problem of declining water quality and decrease in water volume because of rapid urbanization and industrialization. Bangalore city had 262 lakes during the 1960s, presently, just 10 hold water. Similarly, 137 lakes were recorded in Ahmadabad city in 2001, and over 65 stated to be already built over [124]. There are additional examples in India of such a devastation of urban water bodies and deterioration of water quality. Garg et al. [55], have conducted a study on three major lakes of Bhopal viz. Upper Lake, Lower Lake, and Mansarovar Lake in India. After an analysis of the aquatic flora and potential fertility of lake waters, it was discovered that Mansarovar Lake had the

greatest level of eutrophication. Lake Mirik in Darjeeling was studied for its hydro biological properties, and it was seen that in various portions of the lake nutrient concentration very high and potable water quality was spoiled [76]. Other cases of Lake Eutrophication have been reported to Fatheh Sagar lake in Udaipur [36], Sagar Lake in Madhy Pradesh [135], the reasons for deterioration of water quality for all the cases primarily being high nutrient enrichment through anthropogenic activities.

A summary of different cases of eutrophication in major lakes in different parts of the world are summarized in Table 2.1 along with the major causes for eutrophication in the lake. For most of the cases presented here, it has been observed that the major cause of lake eutrophication are human induced ones and, so it is the cultural eutrophication which needs to be given utmost attention. It is evident that cultural eutrophication cases are ever increasing in all parts of the world.

Table 2.1: Lake eutrophication cases from different parts of the world

Region	**Lake**	**Depth**	**The main cause**	**References**
Florida, USA	Okeechobee Lake	Max. 3.7 m, Ave. 2.7 m.	Large *P* inputs from surrounding area.	[162]
Florida, USA	Apopka Lake	Ave. 4.7 m.	High *P* loading from floodplain farms; elevated levels of nutrients, phytoplankton, and suspended matter.	[34]
Louisiana, USA	City Park Lake	-	Sewage contamination from the neighbouring residences	[153]
Washington, USA	17 Western Washington Lakes	-	Internal *P* loading was reported as major cause compared to external sources for summer algal blooms.	[82]
Northern Greece	Pamvotis Lake	Max. 11 m, Ave. 4.5 m.	Discharge from Agricultural, industrial, and urban sectors in last 40 years.	[151, 179]
Poland	Jaroslawieckie Lake	Max. 6.56 m Mean 3.68 m	Large catchment area of agriculture and forests	[136]

Region	Lake	Depth	The main cause	References
Greece	Kastoria Lake	-	Agricultural runoff and seepage from nearby population colonies	[89]
Between Switzerland and Italy	Lugano Lake (Glacial lake)	Max. 288 m Ave. 134 m	High discharge of P (140 mg/m^3) and oxygen in the hypolimnion reduced to zero	[12]
Netherlands	Majority of the lakes	Shallow lakes	The input of P & N from waste water and nearby rivers and canals	[57]
Denmark	Majority of the lakes	Shallow lakes	Higher nutrient load coming from household waste water and agricultural runoffs	[75]
Zimbabwe	Chivero Lake (Ramsar wetland)	Max. 27 m	Hypereutrophic with sewage effluents	[126]
Japan	Biwa Lake	Max. 104 m	Urbanization and industrialization in the vicinity of the lake	[196]
India	Lake Bellandur	Max. 9.21 m	Addition of effluents from the urbanized city and presently in hypereutrophic stage	[24]
Assam, India	Deepor Beel, a RAMSAR wetland	Max. 4 m Ave. 1 m	Continuous flow of sewage and municipal solid waste dumping have degraded the water quality severely	[124]
China	Danjiangkou Reservior	-	Insecticides, herbicides, and organophosphate esters were major potential threats in the reservoir ecosystem	[28]

2.2 BACKGROUND OF LAKE MODELLING

For a successful lake restoration and eutrophication mitigation plan, the basic aim is to identify the sources of pollution and their subsequent remediation. Prevention

of domestic and industrial wastewater disposal into the water bodies, removal of nutrients in wastewater treatment level were used successfully for lake restoration in 1960s in some of the lakes. In 1936, Lake Monona in USA was successfully restored by diverting the sewage to another lake and algal bloom reduced noticeably [41]. But it was obvious that this measure just shifted the problem to the other lake. In a joint effort between US and Canada in 1972, lake restoration had been implemented to Lake Erie which mainly focused on reducing the domestic and industrial input to the lake and it was reported that phosphorus loads from municipal discharges reduced by 84% in 1985 [114]. However, presence of other species such as *Microcystis*, *Anabaena*, *Aphanizomenon* etc. were observed thereafter and other factors such as increased agricultural runoff, climatic factors etc. continued to alter the natural ecosystem balance of the Great Lake [65, 83]. Similarly, water quality improvement in Lake Washington has been reported by Edmondson [41], by nutrient removal in waste water treatment plants.

However more detailed and scientific insight to tackle eutrophication came after the work of Vollenweider [185, 186], where algorithms were provided to calculate nutrient retention in lakes. He gave the concept of permissible levels of total nitrogen (TN) and total phosphorus (TP) by taking the volume of the lake into account such that high volume of nutrient loading is permissible for larger lakes. Other parameters such as internal phosphorus loading from sediments, water retention time, gaseous nitrogen fixation etc. and their role also has been considered in this work. Another major scientific knowledge on eutrophication mitigation was obtained from the work of Schindler [164]. He conducted study on three experimental lakes in Canada for several years and observed that algal biomass is dependent on TP concentration and not on TN as the short-term N deficits are compensated by cyanobacterial fixation of atmospheric N in due course of time. Therefore, it was concluded that it is the TP and not the TN that needs to be removed to prevent eutrophication. But the generality of this concept has been disputed in the later phase by other researchers [16, 48] where mainly N was found responsible for lake eutrophication and considerably low N fixation rate has been observed. Though the debate was centered mainly on N and P, some other micronutrients such as iron, sulphate, silicon, molybdenum, and organic micronutrients (e.g., vitamins) were also paid attention to by Rast and Thornton [145]. But these components were found to be of more importance in the case of marine ecosystem

[149]. Based on these findings several static process-based models have been developed in the succeeding period which were more or less accurate in terms of their predictive power and are site specific. From the late 90's to till date, ecological modelling has developed tremendously. Now it has become possible to find a eutrophication model suitable for any kind of lake with better predictive capacity. But new lake models are still developing day by day, having wider and complex domain for better restoration policy making.

2.3 MATHEMATICAL MODELS

The first eutrophication model was given by Vollenweider in 1968 as discussed earlier [186]. The development of ecological modelling in the succeeding period was based on the scientific insight of this work. The work of Sakamoto [156], where the relation of nutrients and chlorophyll-a was first established by conducting study in some Japanese lakes, were used with data of North American lakes by Dillon and Rigler [39] to form a statistical model between TP and Chlorophyll-a. In this model co-relation coefficient (R^2) was 0.95. Another model that gained popularity in eutrophication control is the OECD model, which is an empirical modification of the Vollenweider model. This model was tested for 87 lakes with varying values of TP concentration and retention time and results yielded co-relation coefficient (r^2) value as 0.86. Many other similar static load models were presented thereafter by Kirchner and Dillon [88]; Larsen and Mercier [93]; Jones and Bachmann [81]; Nicholls and Dillon [127]; Ostrofsky [130]; Chapra and Reckhow [26]; Nurnberg [129]; Ahlgren et al., [2] etc. All these static models were designed in similar fashion. It is considered that the lake is a completely mixed steady state reactor and one or few simple equations were employed to calculate the outflow load of nutrients, mainly P, with outflow water and the quantity of load retained in sediments. The input variables considered in these models are nutrient influx (TP concentration), water retention time and occasionally water mean depth. General output variables are TP in the lake water and TP retention. These models are very simple as the input variable for any lake can be measured and by using simple empirical relationships output variable may be found out. Nutrient concentration in inflow water can be measured directly and by conducting morphometric study, mean depth of water and water retention time can be found out. Due to their simplicity, static mathematical models gained popularity and are frequently used in practical lake

mitigation and control plans. But these models do not account for factors such as internal nutrient loading from the sediments, bioavailable nutrient concentration, seasonal variation of nutrient flux etc. which lead to uncertainty in predicting nutrient concentration for individual lakes [61]. Especially its nutrient predictive efficiency which will seriously hamper in case of recovering lakes having internal nutrient loading where the lakes are not at all in steady state [74]. These factors prompt the development of dynamic mathematical models in recent times and are found to have better accuracy over the static ones.

Mishra [117] has proposed a model for ecosystem dynamics in a eutrophied water body by considering concentration of nutrients, densities of algal population, zooplankton population, detritus, and the concentration of DO as the variables. Here the nutrient input was considered from water runoff from agricultural fields unlike the work of Voinov and Tonkikh [184], where they considered detritus as only nutrient source and not any other external. The model analysis successfully simulated the eutrophication process and was able to draw a relationship between the variables considered. The findings demonstrated that the densities of algae and other aquatic species rise with the increase in the supply of nutrients. It was found that the rise in load of detritus was accompanied with a fall in the dissolved oxygen concentrations. Shukla et al. [169] developed a mathematical model for a eutrophied body of water impacted by organic contaminants. The factors taken into account for the model are the concentration of DO, the concentration of organic pollutants, and the densities of bacteria, algae, nutrients, and detritus. Model analysis revealed that the impact of both water pollution and eutrophication together resulted in decreases the DO level at a higher rate compared to presence of only one phenomenon. Alvarez-Vázquez et al. [7], presented a mathematical eutrophication model to simulate the interplay of nutrient, phytoplankton, zooplankton, organic wastes and DO. For the observed five variables, unique solution has been achieved for eutrophication with non-smooth coefficients.

Malmaeus and Håkanson [115], have given a lake model for forecasting suspended particulate matter (SPM). Various pathways of SPM have been considered such as tributary inflow, indigenous production, sedimentation, transportation, erosion, re-suspension, mixing etc. The dynamic model was tested with empirical data set of several European lakes, and it was seen that it had close agreement with the empirical

data set. All the easily accessible variables such as lake area, mean depth, nutrient inflow, and climatic factors had been considered for model input variables. Based on the model uncertainty and sensitivity analysis, it was concluded that measuring allochthonous and indigenous SPM formation was more difficult to forecast than internal processes such as sedimentation and mineralization.

Yadav et al. [195], have proposed a model for shallow, eutrophic lake Biwa in Japan using satellite remote sensing (a Landsat-8 image) technique. The model was used to evaluate the enclosed area of submerged aquatic vegetation (SAV) and its biomass growth during optimum period. The water clarity was evaluated using a linear regression technique ($R^2 = 0.77$) with the data for periods 2013 to 2016, which was then utilised for SAV categorization and biomass calculation. Water clarity was found to be critical for SAV detection and biomass estimation in shallow eutrophic lakes using satellite remote sensing. The findings demonstrate the practical application of a satellite-based approach for estimating SAV biomass in shallow eutrophic lakes. Huyong Yan et al. [198], have proposed a rough set and multidimensional cloud model (RSMCM) for prediction of trophic status and nutrient status value of water bodies. This model was applied to 24 major lakes and the results were very much consistent and give more accuracy compared to other models.

Zhang et al. [206] have proposed a model for Lake Mogan to show the hysteresis response of aquatic vegetation and water quality with increases phosphate concentration. The dynamic model had a total of nine state variables including P in phytoplankton, zooplankton, the sediment, the pore water in the sediment, the submerged plants, epiphytes, the detritus, soluble reactive phosphate (PO_4^{2-}) and planktivorous fish in the lake. It was found that in between TP concentration 0.16 and 0.25 mg/L, water state changes from fresh to turbid bringing in significant changes in submerged plants. The model was able to predict the shifts from submerged vegetation to phytoplankton at about 0.25 mg/L P-concentration and again from phytoplankton at about 0.10 mg/L P-concentration which was similar to the results obtained by Scheffer et al. [161]. According to the model results, it was concluded that shallow lake restoration process takes place very slowly than eutrophication rate and after reaching threshold concentration of P i.e., 0.25 mg/L, restoration of submerged plants may not be feasible.

Arhonditsis and Brett [9], had proposed a complex multi elemental model for restoration of Lake Washington, USA. The model takes into account multiple nutrient cycles (C, N, P, Si, O), different phytoplankton and zooplankton diversities. The model provided a good fit between the simulated and observed monthly values with less error for parameters like TP, TN, DO, phytoplankton etc. The model could successfully simulate the loading scenario of 1962 (when waste water discharge for the lake was at peak) and was able to forecast the phytoplankton response, nitrogen, and phosphorus cycles accurately. Zhao et al. [209] have proposed consideration of wide range of variables for improvement of eutrophication models. They suggested use of multiple nutrients cycles (P, N, Si, C and O), functional groups of phytoplankton (diatoms, green algae, and cyanobacteria), and inclusion of two zooplankton groups: *copepods* and *cladocerans* into account. Their model results suggest that application of stoichiometric nutrient recycling theory is beneficial in examination of the food quality. The results indicate significant model improvement with the incorporation of multiple nutrient cycle.

Mukherjee et al. [119] have presented a mathematical model for analysis of the carbon cycle in a simulated pond for assessment of cultural eutrophication. The three key activities under the study were photosynthesis, respiration, and decomposition. These processes are crucial for maintaining a dynamic equilibrium and regulating the system's nutritional content. The carbon cycle is typically not taken into account in the previous eutrophication models, but this work demonstrated that the nutrient processes depend on accurate and thorough descriptions of the carbon cycle. Insights from this work have great impact on succeeding eutrophication models. Perhar et. al. [137] have given a zooplankton submodel to study the behavior of zooplankton physiology and behavior by enhancing an existing eutrophication model of Arhonditsis and Brett [10]. The model simulates the interplay of N, P, and highly unsaturated fatty acids through the grazer’s digestive tracks. The model provided good account of different factors responsible for growth of zooplankton under changed trophic status of Lake Washington (USA).

Jensen et. al. [74], have developed an empirical model for seasonal fluctuation of TP concentration to external loading in 16 shallow eutrophic lakes in Denmark based

on 7 to 8 years mass balance data. The model was intended to describe the early recovery phase where external loading is low compared to internal loading. Model results revealed that estimated mean TP concentration deviated 12% on average compared to the observed values and closely resembled the seasonal dynamics after lowering the external loading. However, the model was not able to describe the seasonal variations like transitions from turbid to transparent state. Håkanson and Boulion [62], have developed a model named "Lake Foodweb" to quantify all lake foodweb interactions where a sub-model "LakeMab" was first presented for TP, which is a process based mass balance model. Thereafter this model has been modified by Håkanson and Bryhn [63] for TP applicable to all types of lakes. The structure of the model is basically differential equations based on parameters like mass balance among inflow, outflow, and internal sediment release. Model was tested for different 41 numbers of lakes and it was found that that model could successfully predict the TP concentrations in water. Moreover, it can give additional information regarding TP in sediments, sedimentation, temporal changes, surface, and deep-water volumes etc. not only in surface water level but also in deep water level and sediment level.

Zhang et al. [205] has given a 2-dimensional model to address the problem of eutrophication in Lake Erie, by integrating hydrodynamics with the food web of the lake's lower trophic levels. This fine scale dynamic model considers wind conditions, air temperature, and atmospheric and tributaries related information. This study also considered a through zooplankton model and incorporated the effect of different mussels. A total of 18 types of variables as inputs was considered in the model and validation was done successfully. The findings showed that while the direct grazing consequences of dreissenid mussels on algal biomass are small, the ecosystem is more significantly impacted by the indirect effects of their nutrient excretion.

Seasonal changes in phytoplankton and nutrient budget of Bizerte Lagoon, Tunisia, was reproduced in a 3-D coupled biogeochemical model [15]. Model simulation results showed that plankton production is highest in the inner part of the water body due to high water retention time and contribution of nutrients from boundary. Both N and P loading contribute to the production of plankton. Intensive use of fertilizers in the catchment and contribution of Mediterranean Sea to the Lagoon have impact on the increased plankton growth. Several other authors have presented three

dimensional models to simulate hydrodynamics and nutrient cycling in shallow lakes which are effective in forecasting water quality and algal growth [78, 102, 204].

Page et. al. [131] have conducted a study on the Windermere, Bassenthwaite Lake and Esthwaite Water, in North West England. They have used model PROTECH [45–47] for process representation of algal community using high frequency in lake observations for algal bloom prediction. The model's uncertainty and sensitivity were checked using the Generalized Likelihood Uncertainty Estimation framework. From the model simulations, it was observed that differences between underwater light in real lake systems studied and model's representation and the uncertainties related to nutrients from sediments are of great challenge while forecasting algal bloom. Li-kun et. al. [103], have a proposed a 2-dimensional model for eutrophication in urban lakes of China. The model could describe periodic and region-specific fluctuations in the quality of lake water as well as the interactions between nutrients, phytoplankton, and zooplankton. Comparison between model simulated values and field data revealed that the model was able to estimate the hydrodynamic data and eutrophication processes reasonably with a relative error of less than 11%. The findings indicated that there was a greater concentration of nitrate, phosphate, ammonia, and Chl-a near the lake than in the middle, suggesting that rainwater runoff had a significant role in the algal bloom and that should also be considered.

The hydrodynamic model Hydrax and the water quality model QSim were used by Lindim et al. [104], to study the effect of nutrient reduction (N and P) to lake restoration in a German shallow lake. The simulation results reveal that for a successful water quality control both internal and external P load should be reduced. Reducing both N and P was not found to be beneficial as compared with only P reduction, whereas external N load was not found to have significant impact on ecosystem. It has been reported that P loading of only 400 grams may lead to production of 350 tons of algal bloom. These findings are also in agreement with the findings of Wang and Wang [188]. From the long-term studies in 40 Yangtze lakes and the lakes of North America, it was generalized that N abatement may not decrease the phytoplankton as it can stimulate nitrogen fixing cyanobacteria and P is the key factor governing total phytoplankton biomass. However, effect of climate, nitrogen and other factors should also be considered for proper lake management program [190].

Rucinski et. al. [152] have given a model for forecasting of ecosystem response to altered phosphate loading to Lake Erie. The 1-dimensional model was a linked eutrophication and hydrodynamic model. The model was calibrated with 19 years of data of the lake from the central basin. The results demonstrate that hypoxia can be reduced by controlling nutrient loading, however variations will be observed depending on the climate and meteorological conditions. The model was quite effective in setting phosphorus loading targets to fulfill the water quality standards of the Great lakes. A mass balance model was given by Chapra et. al. [25], to simulate TP values of Lower Great Lakes based on field observed and calculated TP inputs in time series for mitigating water quality in the long run. The model results showed that loading control on Lake Erie would have measurable effect on offshore phosphate concentration in the Great Lakes.

It is quite evident from above discussion that dynamic mathematical modelling has undergone tremendous refinement in recent times and several new processes have been tested for lake eutrophication control specially in cases where adaptation and species composition plays a crucial role in the ecosystem. Coupling of ecological models, hydrodynamic models and watershed models are continuously getting evolved. However, these models necessitate a huge quantity of specific input data, such as hydrological, species composition, meteorological, and bathymetric data, which can be difficult and expensive to obtain, particularly for big and remote lakes. These models frequently rely on oversimplifications and assumptions about the physical, chemical, and biological processes that occur in lakes, which may not adequately depict the complex interactions and feedback mechanisms seen in real-world systems. This can result in considerable uncertainty and restrictions in the model's capacity to replicate the lake's dynamics. Furthermore, these models are computationally demanding and necessitate large computational resources as well as knowledge to set up and run successfully. Finally, calibration and validation of process-based models can be difficult since they necessitate considerable data gathering and comparison with observed data. These factors have led to the popularity of sophisticated machine learning algorithms and advanced uncertainty and sensitivity analysis techniques in lake modelling in recent times.

2.4 DATA-DRIVEN MODELS

Lake eutrophication is a complex process which is dependent on physical, chemical, and biological processes occurring in a lake ecosystem. Traditional mechanistic models, such as nutrient loading models and process-based models as discussed in the preceding section, have been widely used to simulate the eutrophication process in lakes. However, these models have limitations in capturing the complex nonlinear relationships among the variables involved. In recent times, data-driven approaches are being increasingly utilized in lake eutrophication modeling, providing a promising alternative to traditional mechanistic models. Data-driven modelling, also known as empirical modelling, is an approach to modelling that utilizes statistical or machine learning algorithms for analyzing data, pattern recognitions and establishing association among variables. Data-driven models do not explicitly incorporate the underlying physical or biological processes, but instead rely on the relationships observed in the data to make predictions. In recent years, there has been increasing interest in using data-driven modeling approaches in lake eutrophication modeling. This is due in part to the availability of large datasets that can be used to train and validate these models, as well as advances in machine learning algorithms that have made it easier to build complex models. Use of different popular data-driven modelling approaches such as multiple linear regression, artificial neural networks (ANN), adaptive neuro-fuzzy inference system (ANFIS), support vector machines (SVM) and gaussian process regression (GPR) etc. in the field of lake and surface water quality modelling have been discussed in the next sections.

2.4.1 Artificial Neural Network (ANN) Models

Artificial Neural Networks (ANNs) are coming out as a very convenient method for understanding and predicting complicated ecological processes in lake modelling. ANNs can capture nonlinear correlations and interactions between numerous parameters influencing lake dynamics, such as climatic conditions, nutrient loads, hydrological inputs, and biological processes [92, 96]. ANNs may learn and generalize patterns by training on previous data, allowing them to make predictions and simulate future events. Furthermore, ANNs are adaptable to changing environmental conditions and can handle

a wide range of data sources, making them suited for real-time and near real-time predictions. Because of their ability to integrate and analyze enormous amounts of data, ANNs are evolving as a significant method for lake managers, policymakers, and researchers, assisting in decision-making and guiding proactive management plans.

Huo et. al. [69] had used two different ANN model types to predict the water quality of Lake Fuxian in China. Back-propagation and radial basis function neural network was used to model eutrophication indicators TN, SD, DO, and Chl-a. They utilized water quality data of the lake from 2003 to 2008 for the developed models and reported that back-propagation neural networks were efficient in predicting the eutrophication indicators. For all the models, correlation coefficient values greater than 0.7 were achieved and the results of ANN prediction were found better than traditional statistical models.

Karul et. al. [85] have developed a chlorophyll-a (Chl-a) model for predicting Chl-a concentration in Keban Dam reservoir in Turkey using both ANN and linear regression. As the input parameters were quite linearly dependent with Chl-a in the considered dataset, it was found that the prediction accuracy of ANN model were similar to the regression models. In another study carried out by Karul et. al. [84], a three layer Levenberg-Marquardt feed forward neural network model was proposed for Chl-a prediction in three waterbodies in Turkey viz. Keban dam reservoir, Mogan and Eymir lakes. Good correlation was observed between the model predicted and actual values in case of Keban Dam despite the complex ecosystem of the reservoir. For the other two lakes which were smaller in size and were homogeneous, correlation value more than 0.95 was observed between the model predicted and actual values. This study illustrated that non-linear behaviour among ecological parameters could be efficiently modelled with the help of neural networks.

Heddam [68] used feed forward neural network in modelling SD parameter in Saginaw Bay, Lake Huron, USA. Water quality variables dissolved oxygen (DO), total suspended solids, ambient water temperature and chlorophyll-a (Chl-a) collected during the period of 1991 to 1996 were utilized as inputs for predicting SD in Lake Huron. Different input scenarios were considered and the results revealed that total suspended solids and Chl-a were the most influencing parameters whereas DO and water

temperature has less impact in SD prediction. With the neural network approach to predict SD, correlation coefficient of 0.918 was achieved. When compared with multiple linear regression, the developed neural network model showed better prediction accuracy.

Yüzügüllü and Aksoy [202] conducted a study to estimate SD values in Eymir lake in Ankara, with the help of satellite obtained multi-spectral images. For SD prediction, empirical models available in literature was compared with ANN models. Results of the study showed that empirical models performed poorly but with ANN models coefficient of determination (R^2) upto 0.92 was obtained. For the ANN model development different training algorithms were compared and it was found that Levenberg-Marquardt and gradient decent algorithms were better for the considered study. From the results of the investigation, it was reported that ANN is useful tool for getting information from remote sensing data.

Kuo et. al. [92] have proposed conventional back-propagation neural network models to predict DO, TP, SD and Chl-a in eutrophic Te-Chi reservoir in Taiwan considering different input parameters for each model. Using historical data of the lake from 1982 to 1996 for model development, corelation coefficient values greater than 0.7 was achieved for all the models. Similar lake eutrophication models were proposed for prediction of DO, TP, SD and Chl-a using radial basis function neural network (RBFN) approach by Chen and Liu [30] for Mingder reservoir in Taiwan. Two other machine learning approaches in the form of adaptive neurofuzzy inference system (ANFIS) and multiple linear regression (MLR) were also used for prediction of eutrophication indicators. Using linear regression, the input variables of the models were identified. The MLR model results were poor but it was observed that with RBFN and ANFIS models, the non-linear characteristics between input and output parameters can be preserved. Overall, it was reported that the ANFIS model outperformed the RBFN model in terms of prediction accuracy.

Aria et. al. [11] have used conventional multilayer perceptron (MLP) and time delay neural network (TDNN) approach for modelling algal bloom in eutrophic Amirkabir reservoir, Iran. They have used monthly water quality data of twelve years for target prediction. The final input parameters were chosen based on a forward data

trimming method. Results of the study revealed that with MLP models best coefficient of determination (R^2) was 0.93 whereas with TDNN this value was 0.98. It was concluded that though MLP approach is sufficient for immediate estimation of lake water quality status but dynamic TDNN approach may give better accuracy and reliability.

ANN methodology was used to predict Chl-a concentrations in a shallow eutrophic Lake Mikri Prepa in Greece by Hadjisolomou et. al. [59]. Here an optimal k-fold cross validation method was used for training the ANN model as the input dataset was small. Several k values ranging from 3 to 30 were used for model training. It was found that with increase in k number, the prediction accuracy of the ANN model improved, however with leave-one-out cross-validation best results were obtained. The model's sensitivity analysis was performed, and it was reported that with increase in nutrient concentrations (nitrogen and phosphorus) the lake water quality deteriorates further.

Machine learning methods in the form of ANN and support vector regression (SVR) was used by Jimeno-Saez et. al. [77] to model eutrophication indicator Chl-a in the eutrophic Mar Menor coastal lagoon in Spain. Out of nine different water quality parameters, the most important predictors were chosen based on a wrapper feature selection method. Both ANN and SVR models were successful in predicting Chl-a concentrations in the lagoon. All models considering different input combinations showed R^2 value more than 0.7. However, the best result was reported for SVR model having nine input parameters.

A back-propagation neural network was used by Lu et. al. [108] for prediction of Chl-a in Lake Champlain, USA. Monitoring data of 21 years was used to train, test, and validate the model. Based on linear correlation among variables, seven input scenarios consisting of different water quality parameters were used to predict Chl-a. From the findings of the work, it was concluded that ANN model could effectively forecast the Chl-a concentration in Lake Champlain as reflected by higher coefficient of determination value greater than 0.8 achieved during training, validation, and testing phase.

Park et. al. [133] have conducted a study to predict Chl-a concentration in Juam Reservoir and Yeongsan Reservoir in Korea by using ANN and support vector machine (SVM). Meteorological and water quality data of seven years were used for model training and validation. Both the models performed well to predict the temporal variations of Chl-a concentrations based on the weekly data. SVM model's prediction accuracy was found to be better than the ANN model. From the sensitivity analysis results, the connection between environmental factors and Chl-a was identified and effective early warning interval for the considered reservoirs was found to be seven days period.

ANN was used for forecasting DO and BOD values for Gomti River in India by Singh et. al. [172]. The models were trained with ten years monthly data having eleven input water quality parameters. R^2 value greater than 0.7 was achieved for the developed models during training, validation, and testing phase respectively. With sensitivity analysis the major contributing factors for DO and BOD prediction were identified. From the results of the investigation, it was reported that ANN modelling apporoach was effective in predicting water quality variables in the river both spatially and temporally.

Palani et. al. [132] conducted a study to predict dissolved oxygen, chlorophyll-a, salinity, and temperature values both spatially and temporally in Singapore coastal waters. Weekly collected physio-chemical sea water quality data for the period December 1996 to June 1997 was used for model development. ANN training algorithms in the form of MLP and generalized regression neural network (GRNN) were adopted, and it was found that an acceptable correlation was there between actual, and model forecasted values. During model training and testing, R^2 value between 0.8 to 0.9 was observed for all the models. It was reported that despite having uncertainties associated with the sea water quality and dataset limitations, the ANN models were quite accurate and faster in prediction capacity compared to process-based models.

Different machine learning (ML) approaches were used to model phosphate concentration in Feitsui reservoir in Taipei by Latif et. al. [94]. The phosphate concentration in the water body was found to be root cause of eutrophication problem and its fluctuations were non-linear in behavior; ANN, SVM, random forest (RF), and

boosted trees (BT) algorithms were employed for modelling the same. Monthly monitored data of 19 water quality variables from 1986 to 2014 were considered in the study. Based on correlation analysis ammonia, Mg, nitrate, TDS, and hardness were used as model input variables. Based on statistical evaluation of model results it was found that ANN was superior in predicting phosphate concentrations in the reservoir compared to the other ML approaches. Out of different ANN topology tested, it was found that optimum efficiency was obtained for ANN having 3 hidden layers with scaled gradient as training algorithm. R^2 value of 0.979 and root mean squared error value of 1.199 was reported for optimum ANN model in phosphate prediction.

For estimating primary productivity of water bodies, a case study was conducted by Zounemat-Kermani [210] in Hempstead East Marina, New York. Two types of ANN architecture in the form of feed-forward neural network (FFNN) and GRNN were used to estimate concentrations of Chl-a. 14 different physical, biological, and meteorological water quality parameters dataset of 4 years duration were used as model input parameters. Using principal component analysis 8 different groups of uncorrelated variables were found as optimum inputs. The results ANN model performance in predicting Chl-a concentrations were found to be better compared to linear regression models. Out of the two ANN models, the GRNN structure was found to produce more reliable estimation of Chl-a.

Three types of ANN architecture viz. MLP, RBFN, and GRNN along with linear regression model was used by Csábrági et. al. [35] for forecasting DO in the Danube River, Hungary. Using pH, runoff, water temperature, and EC data of the river for the period of 1998 to 2002 as inputs, DO prediction was done for the year 2003. Using R^2, MAE, RMSE, and Willmott's index of agreement as evaluation criteria, accuracy of the ANN models was checked for four different combinations. The prediction accuracy of the ANN models was reported to be better compared to the linear regression model. From the results of sensitivity analysis, pH was reported as most influential parameter in DO estimation.

Sinshaw et. al. [175] conducted a study to determine the TN and TP concentrations during summer season in the US lakes. A FFNN with back-propagation learning algorithm was employed to train the desired models. Regional and national

level water quality dataset from approximately 1000 US lakes were used for the model development. Based on a linear correlation with the outputs three input parameters viz. pH, conductivity and turbidity were chosen as input variables. Results indicated that ANN models were able to forecast the target variables for both regional and national datasets. Models trained with regional datasets showed better accuracy compared to models trained with national datasets. Results of the study revealed that the generalization abilities of ANN models are more in the case of regions having homogeneous climatic and geographical conditions.

A similar FFNN modelling approach was used successfully by Ubah et. al. [182] for predicting water quality indicators pH, electrical conductivity, total dissolved solids and sodium in Ele River Nnewi, Nigeria. Chen et. al. [27] used back-propagation ANN approach successfully to estimate the concentrations of TN, TP, and DO in Chagle River in China. ANN was used by Sarkar and Pandey [158] to estimate DO concentrations of River Yamuna in Mathura region of Uttar Pradesh, India. Considering different input scenarios of monthly water quality dataset of the river for the period 1990 to 1996, high correlation coefficients up to 0.9 were obtained between observed and predicted values. Gazzaz et. al. [56] used a multilayer perceptron network to estimate water quality index (WQI) in Kinta River, Malaysia. Using 23 water quality variables as input and 34 numbers of hidden neurons as optimal structure of the model, a correlation coefficient value of 0.977 was achieved. ANN and partial least square regression method was used to predict DO and BOD levels in Gomti River in India [14] and it was found that ANN outperformed regression methods in output prediction. ANN was successfully used to predict the response of algal blooms against abiotic factors in Kasumigaura Lake, Japan by Wei et. al. [191]. In another case study neural network was used to model total dissolved solids, electrical conductivity, and turbidity by Najah et. al. [122] for the Johar River Basin in Malaysia. It was reported that when used against different data input than the actual training set, the ANN model's efficiency was very robust and authentic.

2.4.2 Adaptive neuro-fuzzy inference system (ANFIS) models

Adaptive Neuro-Fuzzy Inference Systems (ANFIS) is evolving as a powerful tool for addressing the issues associated with understanding and predicting the

complicated dynamics of eutrophication processes in lakes. ANFIS combines the benefits of neural networks and fuzzy logic, allowing expert knowledge and data-driven learning to be integrated [73, 116]. The nonlinear interactions between numerous influencing elements, such as nutrient concentrations, water quality indices, and external loading, that are crucial for eutrophication evaluation, can be successfully captured by ANFIS models. The ANFIS model may learn and adapt to changing lake system conditions by being trained with historical data and adding expert input. ANFIS models are a useful tool for simulating and predicting eutrophication dynamics, allowing decision-makers to have finer knowledge about the effects of nutrient enrichment, and evaluating the success of various management options for eutrophication mitigation. Furthermore, ANFIS models are generally computationally efficient, making them suited for real-time or future applications requiring quick responses for successful lake management. However, accurate and representative data for training and validation, as well as expert input for optimal model design and interpretation is necessary for desired output from ANFIS model.

Mellios et. al. [116] conducted a study to predict the *microcystin* concentration which is a cyanobacterial toxin using water quality parameters as input in ANFIS. The study was carried out in Lake Karla, which is Ramsar site in Greece. DO, Chl-a, TP, nitrogen to phosphorus ratio, and *phycocyanin* were chosen as model input parameters from a 60 valued dataset. In ANFIS training, subtractive clustering method was used and satisfactory prediction accuracy was obtained. In spite of the small dataset the trained model showed R^2 value of 0.73 inferring ANFIS as a promising tool for water quality management.

ANFIS methodology was used for prediction of DO concentration in Gruža Reservoir, Serbia by Rankovic et. al. [142]. Eight water quality parameters gathered monthly over three years period were used in this study and trial and error based approach was employed to find the best combination of input scenario for DO prediction. It was reported that with five input parameters pH, manganese, iron, temperature, and electrical conductivity optimum DO prediction accuracy was achieved.

Another case study utilized ANFIS to predict DO concentrations in Feitsui Reservoir of Taiwan [29]. Here, a conventional ANN and a multilinear regression

model were also used for comparison with ANFIS output. Linear correlation was used to identify the input parameters for DO prediction and pH, conductivity, temperature, turbidity, total hardness, total alkalinity, suspended solids, and ammonium nitrogen were chosen as optimum. In the ANFIS topology, subtractive clustering method was used and 3 numbers of Gaussian membership function was used each input during model training. From the findings of the investigation, it was concluded that the performance of the ANFIS model was better in predicting DO concentrations in the reservoir compared to ANN and conventional linear regression model.

Ahmed and Shah [3] applied ANFIS methodology to predict the biochemical oxygen demand (BOD) of River Surma in Bangladesh. Three years monthly water quality data of ten variables were used for BOD estimation. Two ANFIS models were trained, one with all available parameters (ANFIS-I) and other with parameters having high correlation with BOD (ANFIS-II). Based on statistical evaluation parameters mean squared error, mean absolute error, Nash Sutcliffe efficiency and correlation coefficient, it was observed that both model types were able to predict the BOD values in the river with reasonable accuracy. Comparing both the models it was found that ANFIS-I model was more accurate than the ANFIS-II model.

To check the applicability of ANFIS for prediction of DO concentrations, historical water quality data of Johar River in Malaysia was used by Najah et. al. [121]. Parameters selected for DO prediction were pH, temperature, nitrate, and ammonium nitrogen. Results of the study revealed that the forecasted DO values through ANFIS model were in agreement with the measured values. Sensitivity analysis was carried out and nitrate and ammonium nitrogen were reported as major contributing parameters for DO prediction. The ANFIS model performance was checked against conventional neural network models and the prediction accuracy of ANFIS was found to be superior. However, choosing the best model structure is very important to get the intended efficiency of the model. It was concluded from the study that ANFIS models were quite robust and reliable particularly in case of extreme environmental conditions. These findings were in agreement with the work of El-Shafie et. al. [42] where ANFIS was used to inflow forecast in case of Nile River to the Lake Nasser reservoir.

Al-Mukhtar and Al-Yaseen [5] conducted a study to estimate total dissolved solids (TDS) and electrical conductivity (EC) of water in Abu-Ziriq marsh in Iraq using ANFIS, ANN and linear regression models. Monthly water quality data for the period 2009 to 2018 was used for the study. Based on correlation values input parameters were determined for the desired outputs. Based on statistical evaluation parameters correlation coefficient, root mean square error and Nash–Sutcliffe efficiency coefficient, it was observed that ANFIS model results were more accurate compared to that of ANN and regression models.

Bruder et. al. [17] studied about the application of ANFIS in prediction of cyanobacterial metabolites which are taste and odour producing compounds in algal bloom affected Eagle Creek Reservoir in USA. ANFIS was used to understand the complex quantitative relationship between various physio-chemical variables, algal species, and metabolites. ANFIS models were developed for metabolites geosmin and 2-methyl isoborneol using observations from 2008 to 2010 on the reservoir. It was found that ANFIS models were able to predict the geosmin and 2-methyl isoborneol with acceptable accuracy and R^2 value of 0.83 and 0.82 was reported for the models respectively.

ANFIS methodology was used by Luo et. al. [110] for prediction of Chl-a concentrations in US lakes. In this study water quality data collected from 2007 to 2012 for more than 1000 lakes in US was used for Chl-a prediction. It was seen from statistical analysis that man-made or artificial lake's water quality parameters were associated with more bias and uncertainty compared to the natural lakes where strong correlation between water quality variables was observed. TP, TN, turbidity, and SD were used as predictors for estimation of Chl-a and different ANFIS topology was tested and compared with multilayer perceptron neural networks. In case of ANFIS models, fuzzy c-mean clustering method, grid partition method and subtractive clustering methods were used to model Chl-a. Results of the study showed that ANFIS model with fuzzy c-mean clustering method was best for prediction of Chl-a in natural lakes but for man-made lakes result of neural network model was more promising.

In an another case study ANFIS was used to model DO concentrations for Klamath River in USA by Heddam [67]. pH, temperature, sensor depth and

conductance were the input variables for DO prediction and both grid partitioning method and subtractive clustering method were used for ANFIS model development. Both ANFIS model were able to predict the DO concentrations precisely and coefficient of correlation was reported between 0.992 to 0.998 during training and testing stages. Emamgholizadeh et. al. [49] conducted a similar study to model DO, chemical oxygen demand (COD) and BOD for the Karoon River in Iran. They have used ANFIS, neural network in the form of a multilayer perceptron (MLP) and radial basis function was used as modelling tools. Dataset comprising monthly values of water quality variables for 17 years was considered in the study and 9 inputs were used for each model. From statistical analysis of model results it was observed that both ANFIS and ANN models were able to rationally predict the DO, COD and BOD values in the river, but performance of the MLP models was found superior to the ANFIS models.

Pham et. al. [138] conducted a study to forecast water quality index (WQI) using ANFIS, ANN and group method of data handling (GMDH) techniques for the free surface wetland constructed in University Sains Malaysia. Different water quality parameters such as pH, phosphate, nitrate, nitrite, DO, BOD, COD, suspended solids, conductivity, and ammoniacal nitrogen monitored for 14 months duration were used as model data base. It was reported that ANFIS model with Nash-Sutcliffe Efficiency 0.9634 and mean absolute error 0.0219 provided better prediction accuracy for WQI than the ANN and GMDH approaches. Findings of sensitivity analysis with ANFIS revealed pH, ammoniacal nitrogen, suspended solids, and COD as the major influencing parameters for WQI prediction. Similar modelling approach was undertaken to model WQI for Selangor River in Malaysia by Yaseen et. al. [201]. They have used different ANFIS structures based on fuzzy c-means clustering, subtractive clustering and grid partitioning to predict WQI and results of the study provided a good correlation of model simulated and actual values. From both the studies it was concluded that soft computational tools like ANFIS provide an upper hand in ecological modelling in comparison with conventional models that are costly, laborious, time consuming and associated with uncertainty.

ANFIS was used to classify river water quality status of major rivers in China by Yan et. al. [197]. Water quality parameters DO, COD and ammoniacal nitrogen collected weekly for nine weeks period from all the major river basins in China was

used for classification. Eight ANFIS models with various membership functions were trained and it was found that Gaussian membership function provided optimum prediction accuracy. In an another case study ANFIS along with SVR, multi-layer perceptron (MLP) and radial basis function neural network was used for flow forecasting at the Buna River in Eastern Herzegovina, Bosnia and Herzegovina [90]. Using RMSE and MAE as evaluation criteria, ANFIS model was found to outperform other methods in short-term river flow prediction.

Elkiran et. al. [44] had applied different machine learning techniques for single and multi-step ahead predictions of DO in Yamuna River, India. ANFIS, ANN, SVM, linear Auto regressive Integrated Moving Average (ARIMA) and a few ensemble methods were employed for target prediction. Monthly monitored data for parameters BOD, COD, discharge, ammonia, water temperature, and pH from the year 1999 to 2012 were used for DO forecasting. The performance of the models were analysed based on R^2 and RMSE values and ANFIS model's accuracy was reported as more reliable compared to other algorithms. From the findings of the work, it was also reported that introduction of ensemble techniques with ML algorithms like ANN or ANFIS could lead to more robust modelling efficiency of water quality parameters.

For estimating water level fluctuations of Lake Beysehir in Turkey, machine learning approaches in the form of ANFIS, ANN and seasonal autoregressive integrated moving average (SARIMA) was used by Yarar et. al. [200]. The models were trained using rainfall, water level and evaporation data of the lake collected between 1966 to 1984. All the models showed promising results but maximum accuracy in prediction was obtained for ANFIS model where minimum mean squared error value of 0.0057 and maximum R^2 value 0.793 was reported. In another study different data driven approaches were used to model lake water level fluctuations for the Manyas and Tuz Lake in Turkey by Sanikhani et. al. [157]. ANFIS methodology with grid partitioning and subtractive clustering were used to predict one to three months ahead level fluctuations in the considered lakes and it was concluded from the investigation that ANFIS is a useful tool for lake water level forecasting. Similar ANFIS approach was successfully used for water level fluctuations prediction in Lake Urmia in Iran [181].

Ly et. al. [111] conducted a study for eutrophication prediction using machine learning (ML) algorithms for urban Han River in South Korea. 20 water quality parameters monitored on monthly basis for a period of 2011 to 2020 was used for algal bloom prediction using 8 different ML techniques. TP, TN, pH, BOD, temperature, DO, precipitation, and flowrate were identified as critical parameters for algal bloom were used as input variables in the considered ML networks. Out of all the models tested it was seen that ANFIS model results were better both qualitatively and quantitively compared to other methods.

2.4.3 Use of SVM and GPR in water quality modelling

To address the concerns related with understanding and predicting water quality of surface water bodies, Support Vector Machines (SVM) and Gaussian Process Regression (GPR) have been used in ecological modelling. SVM is a machine learning technique that can successfully capture nonlinear relationships and works well with high-dimensional data. It may be trained to categorise and predict a lake's eutrophication status using historical data on nutrient concentrations, water quality measurements, and other relevant variables. SVM has been effectively used in eutrophication modelling to categorise lakes into different trophic stages and forecast the risk of eutrophication.

Gaussian Process Regression (GPR), on the other hand, is a probabilistic model that captures errors in predictions and provides a flexible framework for water quality modelling. GPR has been utilised in eutrophication modelling to determine the effect of external factors on lake water quality, like nutrient loading, climate change, etc. It is capable of producing probabilistic projections of multiple water quality indicators, allowing for a more comprehensive knowledge of the eutrophication process and associated uncertainties. SVM and GPR have both demonstrated potential in water quality models by delivering accurate forecasts and significant insights. It should be noted that the performance of these models is dependent on the availability of high-quality data, adequate feature selection, and thorough model calibration. Expert expertise and domain-specific information are also essential for interpreting results and incorporating ecological awareness into the modelling process.

Riza et. al. [123] conducted a study for monitoring and prediction the water quality of Langat River in Malaysia by application of several machine learning applications. For prediction of total suspended solids, total solids, and dissolved solids in the river 4 regression tree models, exponential and rational quadratic GPR, 2 SVM models fine and medium Gaussian, and traditional ANN methods was tested. 25 different water quality variables were used for target prediction. From the results of the study, it was seen that both GPR and SVM models were not able to predict the target variables with desired accuracy and problem of overfitting was observed which was attributed to smaller input dataset. ANN models were able to correlate the predicted and observed values of the river water and outperformed other models.

For early assessment of eutrophication in lakes, Garcia-Nieto et. al. [53] had used a hybrid algorithm to model Chla-a and TP concentration in Pozon de la Dolores Lake, Spain. SVM approach coupled with particle swarm optimization technique was used to predict eutrophication indicators based on physio-chemical and biological input variables. It was found that the hybrid model could efficiently predict the eutrophication indices and R^2 values of 0.90 and 0.92 was attained for the TP and Chl-a model respectively. Garcia-Nieto et. al. [54] had conducted a similar study to model Chl-a and TP concentrations for eutrophication management in the Englishmen Lake, Spain. SVM integrated with artificial bee colony, M5 model tree and a MLP neural network was used for modelling of eutrophication indicators. Results of the study showed that M5 and MLP model results were not promising however very good prediction accuracy was obtained for the SVM model for TP and Chl-a prediction. The same authors had investigated the applicability of GPR in eutrophication prediction by using ten years water quality dataset of Tanes Reservoir in Spain [52]. Using seven biological and thirteen physio-chemical variables as inputs, GPR integrated with Limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGSB) method as an optimizer was used to model the Chl-a concentration in the reservoir. It was observed that a satisfactory correlation between model predicted and actual values were achieved through GPR based modelling as reflected by R^2 and correlation coefficient values of 0.8597 and 0.9306 respectively.

Different machine learning algorithms were used by Nouraki et. al. [128] for prediction of water quality parameters in Karun River, Iran. SVR, MLR, M5 model tree,

and random forest regression (RFR) techniques were used for prediction of total dissolved solids (TDS), total hardness (TH) and sodium absorption ratio (SAR) in the river system. Using principal component analysis (PCA), the number of inputs were optimized for each target output. It was found except M5 models other models could successfully predict the observed values of the target values. From the results of the study, it was concluded that such machine learning approaches can provide a better understanding of ecological systems compared to other mathematical physical process-based models.

Singh et. al. [173] have used SVM for prediction of BOD for the Gomti River in India based on monthly water quality data of 13 variables collected over a period of 15 years. Based on a linear regression approach the input feature selection was done and pH, total alkalinity, phosphate, COD, DO, and chloride were found suitable for BOD estimation. The SVR model was able to predict the BOD values with reasonable accuracy and correlation coefficient values of 0.952, 0.909, and 0.907 were reported for model training, validation, and testing respectively. SVR, ANN and regression modelling approaches were used by Alnuwaiser et. al. [6] for estimating river water quality in terms of total dissolved solids (TDS) and electrical conductivity (EC). Seven water quality parameters were used for target prediction and performance of the models were compared. R^2 value of 0.97 was observed between model predicted and observed values for SVR model and relative performance of SVR was found better than ANN and regression models.

For eutrophication management in shallow eutrophic Chaohu Lake in China, SVR was used to formulate predictive models for Chl-a, TN, and TP by Xu et. al. [194]. Water quality data comprising of monthly measured values of parameters pH, temperature, time, velocity of flow, disturbance, Chl-a, TN, TP, and DO values collected from the year 2000 to 2008 in the lake were employed for model development. During SVR model optimum architecture selection a five-fold cross-validation was used and tradeoff constant (C), insensitive loss function (ε), and kernel function was fixed by using a trial and error estimation. Using a radial basis kernel function, and optimum values of C and ε, SVR models for Chl-a, TN, and TP produced correlation coefficient values of 0.976, 0.993, and 0.996 respectively. Comparing the results of SVR models for eutrophication indicators with conventional back propagation neural networks, it

was concluded that SVR models were able to predict the target indicators with more precision.

Luo et. al. [109] conducted a study to establish relationship model between Chl-a and ecological variables with the help of SVR in Three Gorges Reservoir in China during spring algal bloom. Water samples collected weekly from 2007 to 2008 were analysed and based on a stepwise linear regression method DO, pH, phosphate, turbidity, ammonia, nitrate, and silica were used as model independent variables. C and ε were optimized with leave one out cross-validation technique and optimal SVR model architecture was finalised. R^2 value of 0.820 was obtained for SVR model and it was found that the results were superior compared with MLR and ANN model for Chl-a prediction in the reservoir.

Fan et. al. [51] had conducted a study to illustrate the applicability of SVM in predicting bio-indicators in the Taizi River, China. Using common physio-chemical and hydro morphological parameters of the river as inputs, SVM models with 10-fold cross-validation technique were developed for prediction of fish, macroinvertebrates, and algal communities. The SVM model hyper-parameters were optimized using genetic algorithm (GA) and it was found that developed models were able to yield good prediction accuracy of biological indicators.

Machine learning algorithms SVM, ANN and group method of data handling (GMDH) were used for forecasting water quality variables for Tireh River in Iran [60]. SVM model with radial basis kernel and tansig as transfer function in ANN was used to model nine water quality parameters based on the data collected over a period of 55 years. Comparing the results of model prediction accuracy SVM and ANN showed good correlation between observed and predicted values. The GMDH models resulted in lowest accuracy. Overall SVM models were found to produce less error in estimation and so were recommended over the other methods.

GPR was used for sulphate content prediction in the lakes of China by Zhao et. al. [208]. Different water quality parameters such as pH, transparency, temperature, DO, conductivity, Chl-a, TP, TN, and ammonia-N collected from Yangtze River basin during the period 2007 to 2009 were used as predictors for sulphate estimation. In GPR

topology different kernels were tested and exponential kernel with 5 fold cross-validation method was found to produce least amount of error. The GPR was compared with other data-driven methods such as linear regression, SVR and decision trees and it was reported that GPR produced better prediction accuracy for sulphate estimation in lakes.

Suphawan and Chaisee [180] had demonstrated application of GPR along with ANN and MLR in predicting water quality index (WQI). Climatic variables such as monthly average values of temperature, rainfall, evaporation, and humidity were used for prediction of WQI in Ping River basin, Thailand. The findings of the research demonstrated that the three methods were able to predict the WQI with acceptable accuracy, GPR simulations were found to be more robust and accurate. Rainfall parameter was reported as the most significant one for prediction of WQI for the studied river.

Zara Farjoudi and Alizadeh [203] used GPR to understand the relationship and prediction of total dissolved solids (TDS) with anions and cations present in river water. 16 year's monthly data of anions and cations in Iran's Tajan River water were used for TDS estimation. Five kernel functions were used for GPR modelling and their parameters were finalized by a trial and error approach. It was found that GPR model with rational quadratic kernel provided good performance and R^2 value of 0.9836 was observed between model predicted and actual values of TDS. The performance of the GPR model was compared with ANN model for TDS prediction. The TDS values predicted by GPR model were found to be better than the ANN predicted values. In another study TP, TN, and DO concentrations in surface water bodies Songhua River and Liao River in China was modelled successfully with GPR by Liu et. al. [105]. Apart from lake and surface water quality prediction, GPR modelling was also found to be effective tool for estimation of ground water level fluctuations in drought prone areas [43].

2.5 CONCLUDING REMARKS

From the above discussion it is quite evident that anthropogenic activities have resulted in substantial growth of lake eutrophication cases all over the world. For

restoration and management of eutrophied lakes several modelling approaches have been explored in the last four to five decades and they are still evolving. Mathematical and process-based models have been used extensively in lake modelling and such models can simulate physical, chemical, and biological processes in lakes giving significant insights into the underlying mechanisms causing eutrophication. These models include complex factors such as nutrient loading, hydrodynamics, and ecological interactions to simulate the growth and dynamics of algae and other aquatic organisms. However, process-based models rely on extensive input data, expertise, and computational resources, and often based on simplifications and assumptions which may affect their accuracy.

Lake eutrophication modelling has seen a significant rise in popularity of data-driven and machine learning approaches. The intricate links between nutrient loading, water quality metrics, and eutrophication dynamics have been captured using techniques like Support Vector Machines (SVM), Gaussian Process Regression (GPR), adaptive neuro-fuzzy inference systems (ANFIS), and Artificial Neural Networks (ANN). Based on historical water quality data real time and future predictions can be made. These models provide benefits including flexibility, handling high-dimensional data, and the capacity to identify nonlinear patterns. They have demonstrated potential in forecasting eutrophication indicators, trophic states, identifying crucial water quality indicators responsible for eutrophication, and evaluating the efficiency of management policies. In majority of the cases for eutrophication prediction, water quality parameters such as Chl-a, DO, TP, TN, SD etc. were used as predictor variables. Data-driven models are quite sensitive to spatial and temporal variations and are generally case specific, however few research suggested that generalization can be done for models under homogeneous environmental and geographical conditions.

Major concern regarding data-driven techniques still remains as difficulties with data availability, model interpretability, and model uncertainty assessment. Data availability is of utmost importance in machine learning models. The quality, quantity, and representativeness of the data directly impact the performance and reliability of the models. It is seen that almost all the case specific lake eutrophication models using data-driven approach as discussed earlier were developed with water quality parameter datasets collected over a long duration of time like two to twenty years by a concerned

monitoring authority which is evident from Table 2.2. 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 with data-driven modelling approach becomes a concern for stakeholders and policy makers.

Table 2.2: Application of data-driven modelling to predict lake eutrophication

References	Location	Eutrophication indicators	Modelling approach	Data collection duration
[84, 85]	Keban Dam reservoir, Mogan Lake, Eymir Lake in Turkey	Chl-a	ANN	1991-1996
[92]	Te-Chi Reservoir, Taiwan	DO, TP, SD, Chl-a	ANN	1983-1999
[4]	Omerli Lake, Turkey	DO	ANN, MLR	1990-2004
[71]	Lake Fuxian, China	DO, TN, SD, Chl-a	ANN	2003-2008
[29]	Feitsui Reservoir, Taiwan	DO	ANN, ANFIS, MLR	1993-2011
[30]	Mingder Reservoir, Taiwan	DO, TP, SD, Chl-a	ANN, ANFIS, MLR	1993-2013
[68]	Saginaw Bay, Lake Huron, USA	SD	ANN, MLR	1991-1996
[54]	Englishmen Lake, Spain	Chl-a, TP	SVM, M5 tree, ANN	2006-2014
[11]	Amirkabir Reservoir, Iran	Bacillariophyceae species	ANN	2000-2012
[194]	Chaohu Lake, China	Chl-a, TN, TP	SVR	2000-2008
[52]	Tanes Reservoir, Spain	Chl-a	GPR	2006-2015

* ANN= Artificial neural network, ANFIS= Adaptive neuro-fuzzy inference system, MLR= Multiple linear regression, Chl-a= Chlorophyll-a, DO= Dissolved oxygen, SVM/SVR= Support vector machine/regression, TP= Total phosphorus, TN= Total nitrogen, SD= Secchi depth, GPR= Gaussian process regression