ENVIRONMENTAL PROBLEMS

Vol. 3, No. 2, 2018

# MATHEMATICAL SIMULATION FOR ALGAL GROWTH IN THE WATER RESERVOIRS OF MONCTON CITY (NEW BRUNSWICK, CANADA) BY THE SUPERVISED LEARNING CLASSIFIER

**Qurat-Ul An Sabir, Muhammad Nadeem, Tri Nguyen-Quang**

*Biofluids and Biosystems Modeling Lab (BBML)*
*Department of Engineering, Faculty of Agriculture, Dalhousie University*
*39, Cox Road, Truro-Bible Hill, B2N 5E3, Canada*
*tri.nguyen-quang@dal.ca*

*Received:*



**Abstract.** Mathematical model is a good approach to deal with the coupling effects of governing parameters in algal bloom growth. Among many models to deal with combining factors and data-based supervised learning classifiers, the Artificial Neural Network (ANN) has the most significant impact on the development of bloom pattern. The objective of this paper is to use the Artificial Neural Network (ANN) model to simulate the growth of harmful algae under environmental factors that can lead to bloom pattern in two reservoirs of Moncton city (Canada) with the collected data from two years of observation 2016–2017.

**Key words:** Artificial Neural Network (ANN), Cyanobacteria, Harmful Algal Blooms (HAB), Modified Redfield Ratio (MRR), Supervised learning classifier

## 1. Introduction

Field experiments are extremely necessary to elucidate various factors that affect algal blooms and their proliferation. However, parameters that can be determined in field experiments are limited due to their cost and time-consuming abilities, especially as you can never measure enough 'all parameters' involving in the pattern for comparison. Moreover, their combining effects on the algal growth play a significant role more than each single factor. Therefore, the mathematical model is a good approach to deal with the coupling effects of governing parameters in the bloom occurrence and proliferation. We expect to study the key factors which govern the algal dynamics and then establish an algal response model which could effectively predict algal blooms.

The model we want to use for this paper is the type of supervised learning classifiers, named Artificial Neural Network (ANN), which has already been developed as a model utilizing the learning process algorithms and considered as an alternative tool to data processing in ecological sciences. An important step in the development of ANN is to select the input data which have the most significant impact on the model's performance (ASCE, 2000). Models for HAB involves the development of a functional relationship between the algal growth and environmental parameters. The ability to ANN technique to capture the behavior of non-linear complex data (Elangasinghe et al., 2014). The models type ANN are now being widely used because they can learn (training), non-linear and input relationship by using data and parameters. ANN has many characteristics like generalization, adaptation, etc. and multilayer model trained with backward propagation is the most used in modeling and optimization, prediction, function approximation (Madic & Radovanovic, 2011).

The objective of this paper is to use the mathematical model Artificial Neural Network (ANN) to simulate the growth of algae under the environmental factors in two reservoirs in New Brunswick province of Canada: Irishtown (for recreational water) and McLaughlin (for back-up drinking water) with the collected data from two years of observation 2016-2017. Our study is based on a set of five main parameters for the input dataset: 1) Total Phosphorus (TP); 2) Nitrates; 3) Temperature; 4) pH and 5) Dissolved Oxygen (DO) and two outputs in this first step model: Chlorophyl-a (Chl-a) and Phycocynin (PC), representing the growth of harmful algae.

## 2. Methodology, Mathematical model and Analysis process

### 2.1. Study sites and data collection

In the city of Moncton (New Brunswick, Canada) there are two reservoirs that are experiencing toxic algal blooms over the last few years. McLaughlin Reservoir (McL) is the city's back up drinking water supply and Irishtown Reservoir (IR) is a recreational reservoir, part of the Irishtown Nature Park. These reservoirs have been experiencing algal blooms since 2009.

Sampling locations were already predetermined plus some other points having noticeable patterns of blooms or scums that we decide to take during our field trip observation. The YSI probe (Professional Plus, Hoskin scientific LTD, USA) was used for Dissolved Oxygen (DO), conductivity, temperature and pH at each sampling point. As soon as sampling was finished, they were taken to the lab for micronutrient analyses as well as PC and Chl-a measurements by Fluorometer, and for taxonomy and toxicity. In the context of this paper, taxonomy and toxicity will be not mentioned herein.

These two reservoirs, although they are in the same region of Moncton city and not very far from each other (16km), can be considered as independent due to the fact that one (Irishtown) is served for the public recreational waterbody and another one (McLaughlin) is strictly not accessible to public, used for the back-up drinking water of the municipality.

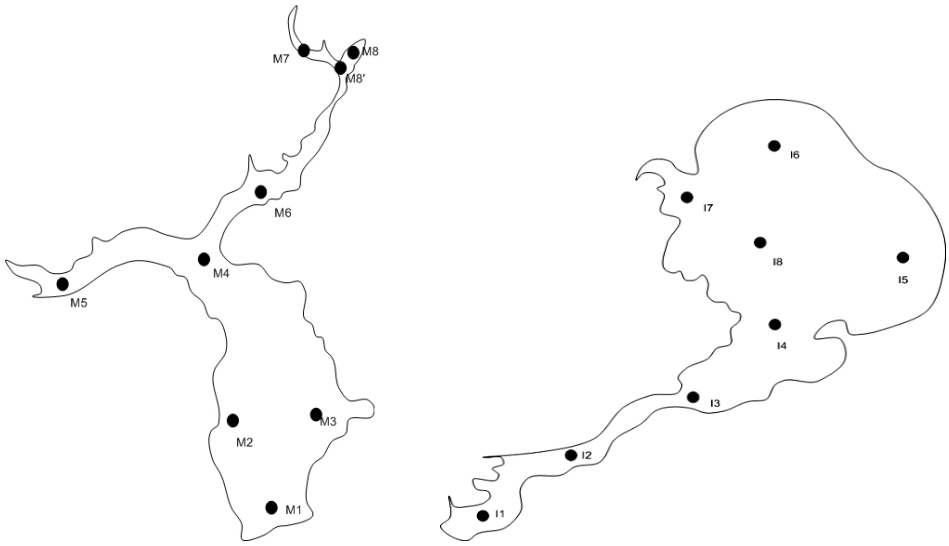


**Fig. 1**. Sampling locations for McLaughlin Reservoir (left) and Irishtown Reservoir (right) in 2017

## 2.2. Conception and configuration of Artificial Neural Network Model (ANN)

Inspired by the biological nervous system and artificial intelligence (McCulloch and Pitts, 1943; Bishop, 1994), the ANN model is applied in our paper consists of many simple processing elements that are variously called neurons or nodes. Each neuron or node is connected to other nodes by means of direct communication links, each with an associated weight function. The weights represent information being used the net to solve the problem. We suggested the back-propagation multi-layer neural network which is the most common and convenient ANN used many research works. In this network, the nodes are arranged into 3 categories of layers: input layer (Observations), hidden layer(s) (intermediate nodes) and output layer(simulated results) (Fig. 2*a*, 2*b*).

The input layer receives information from the outside world, processes it and then transmits it to the hidden layers and then to the outcomes (output layer). Predictions of the problem will be communicated via the output layer. Each node in hidden layer receives input signals from multiple inputs in proportion to their connected weight functions.

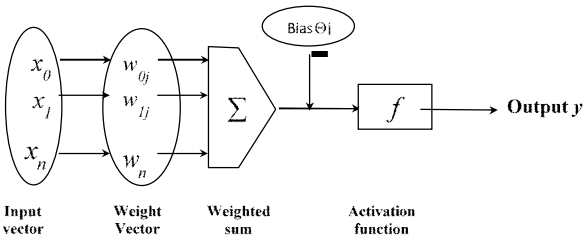


**Fig. 2a.** General conception of the ANN model (The inputs to unit j are outputs from the previous layer. These are multiplied by their corresponding weights to form a weighted sum, which is added to the bias associated with unit j. A nonlinear activation function f is applied to the net input)

A nonlinear activation function f is applied to the net input)

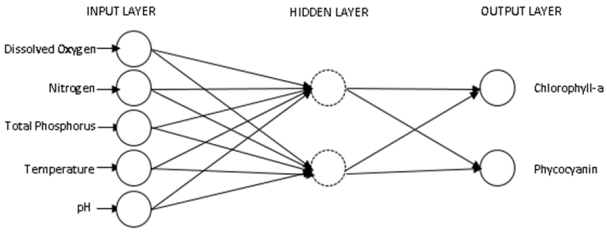


**Fig. 2b.** Our suggested ANN model

Each connecting link to each node is characterized by a weight W of its own. The weights on the

connections from the input $X_1$, $X_2$, … $X_n$ to the neuron Y are $W_1$, $W_2$, …$W_n$, respectively.

An 'adder' (term used by Huang and Foo, 2002) for summing all the weight signals from input layer. This adder is explicitly expressed by a linear combiner 'V'.

$$V = W_1X_1 + W_2X_2 + \ldots + W_nX_n \quad (1)$$

Transfer function F for liming the amplitude of the output of a node. With this function, the output signal can be expressed by

$$Y = F(V) \quad (2)$$

The nonlinear functions are commonly used for most of the ANN model under the form of sigmoid or tangent hyperbolic. When the model is a multilayer neural network (i.e. the network architecture comprises input layer, hidden layer(s), and output layer, as illustrated in Fig. 2*a*), the weight signals from one layer to other one takes the form of the weight matrix $W_{ij}$. The iterative process of adjusting the network connection weights in response to a few examples presented to the network is called ''training''. The basic idea of training is to achieve a unique set of connection weights needed to calculate outputs that are very close to the expected outputs for all the examples used in training. The trained neural network can be then used to predict outputs corresponding to a set of new inputs. A sufficiently trained network is expected to produce results that satisfy the actual outputs.

A back-propagated artificial neural network (BP-ANN), multiple layers feed forward, learning system with a mathematical function and Levenberg-Marquard learning algorithm was used for model development. The difference between BP-ANN and other ANN algorithms is the way in which weights are adjusted for accurate predictions (Maier and Dandy, 2000). The equation 3 below explains how the input nodes are converted into output (H) using a transfer function. This mechanism is repeated for all preceding nodes in a network till the final layer is achieved. The training of the architecture involves a mechanism of providing the network with the desired output with the efficient network performance. The network will estimate the output value from the inputs, compares the model predicted output to the target value, and then adjusts the weights in order to reduce the errors between the network output and the target values. The network training is achieved if the error is less than a given value called 'tolerance'. The error minimisation process is achieved by using the gradient descent method (Bishop, 1994).

$$H = I_1W_1 + I_2W_2 + I_3W_3 + \ldots\ldots I_nW_n \quad (3)$$

Where

*I = Inputs; W = Weight function; H = Output; n = 1 to 5.*

The size of the input layer neurons corresponded to the number of input parameters with one hidden layer which is considered enough to model majority of continuous non-linear function. More hidden layers may cause over and under fitting of the network (Torrecilla et al., 2004). In summary, the neural network performs a non-linear transformation on the input variables (X) to achieve an output (Y). This phenomenon is explained in equation 4.

$$\{Y\} = f(\{X\}) \quad (4)$$

Where

*Y = Output; f = Non-linear function; and X = Input variables.*

Our suggested configuration for ANN model is illustrated in Fig.2b with three main parameters for the input layer: 1) Total Phosphorus (TP); 2) Nitrates; 3) Dissolved Oxygen*;* 4) *pH and* 5) *Dissolved Oxygen (DO)* represented by 5 variables $X_i$= 1 to 5 (5 input nodes). The two outputs are the quantity of Chl-a and PC standing for the growth of cyanobacteria.

The variability in the training and validation datasets was examined through the summary statistics (minimum, maximum, mean value, standard deviation and coefficient of variation).

## 2.3. ANN software and simulation process

A commercial software, Peltarion Synapse (Peltarion Systems®, Netherlands, with license) was utilized to develop the ANN model. This software allows the user to define the architecture of the network and offers a variety of weight functions and having an ability to manage the critical network parameters such as epoch size (time step of calculation) in terms of the mean square error (MSE), root mean square error (RMSE), and coefficient of correlation $R^2$ (this coefficient is defined as the goodness of the simulation curve approximating to the real data points. An $R^2$ tends to 1 indicating that simulation curve fitting well with the data).

We processed herein nine different architectures for our hidden layers and tested them in order to find out a suitable design which can fit with data, and hence for the prediction of algal bloom development. The training step was processed with various hidden layers/nodes and weigh functions. We started with the simplest scenario and moved to the more complex one. Five mathematical functions (Tanh sigmoid, Linear, Exponential, Morlet and Logistic Sigmoid) were tested for each architecture. Many simulations were performed to choose the best configuration of our ANN model (optimal design of the architecture including input, output and hidden layers, best mathematical function for weights) based on the selection criteria $R^2$, MSE, RMSE of each simulation run.

Regarding the time step for simulation (which is called hereafter 'epoch' by Peltarion Synapse), to determine its optimal value, the best selected ANN configuration will be operated at different epoch values with an interval of 1000, and the values for MSE and RMSE are recorded at each

interval. The epoch values are plotted against MSE and RMSE to find the optimum epoch for the ANN configuration. Time step (or epoch size) has been demonstrated to have a major influence on the error terms.

Once our best configuration is structured and trained, the performance of developed ANN model will be validated internally and externally. The performance of the developed model in predicting Chl-a & PC using an independent dataset was tested and evaluated in terms of $R^2$, MSE, RMSE. The operational protocol of the developed ANN model including training and validating steps is presented in Fig. 3.

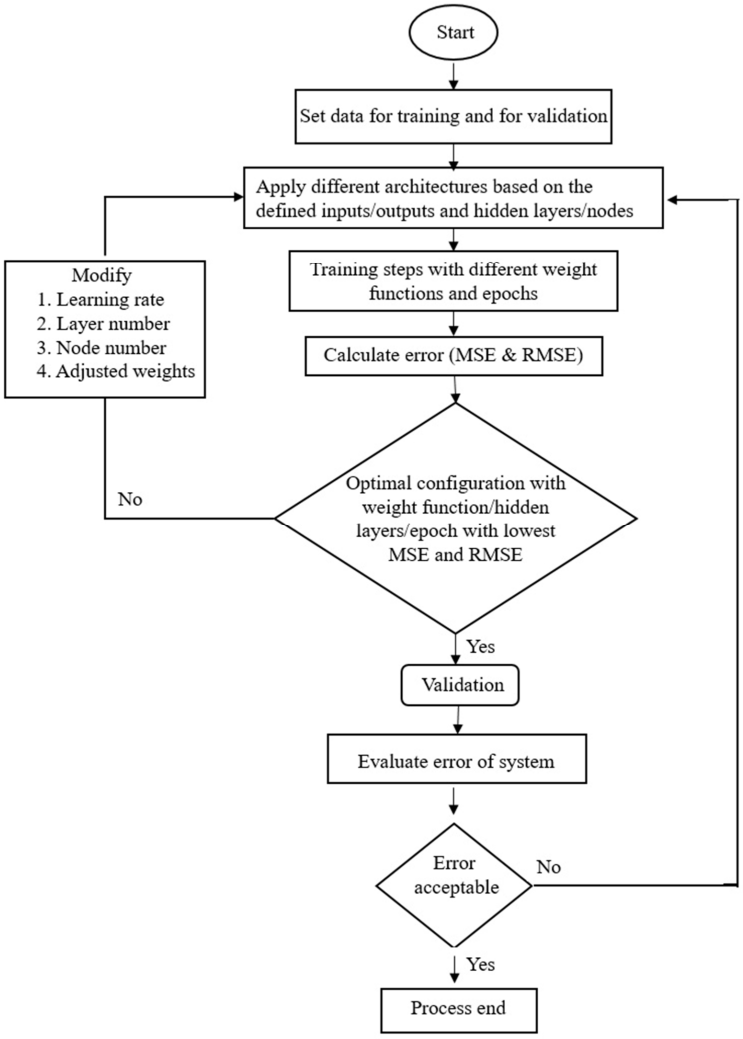


**Fig. 3.** Algorithm for the ANN model development

When we obtain our resulting outputs by model, we can compare first with the experimental data and once they are checked, we can process the model towards the prediction step.

The validity of the model assumptions (constant variance of the error terms) was tested by examining the residuals at 5 % level of significance. Independence of error terms was assumed to be valid through the randomization of the treatment combinations.

## 2.4. Data normalization

One of the key step for data preparation to enhance the model performance is the normalization of dataset, and hence outputs obtained were also normalized quantities. We must convert back the normalized output quantities to the real values of PC and Chl-a. The following formula (5) is used for the normalization process:

$$u_i = \frac{(R_i - Min_i)}{(Max_i - Min_i)} \quad (5)$$

Where

- $u_i$ = Normalized value of input; $R_i$ = Actual value of input; $Min_i$ = Minimum value of input.
- $Max_i$ = Maximum value of input.

The normalization is necessary for the quick convergence of training process as well as for an accuracy of selected weight functions, since normalized

quantities yielding into the range [0,1] hence much easier to be controlled.

## 3. Results and discussions

### 3.1. Choice of an ANN optimal configuration for the predictive model

Table 1a, b presents the statistical summary of the selected data for our simulation. After the determination of a best fitting weight function, all developed architectures were tested and evaluated with it to configure the optimal setting of the network (number of hidden layers, epoch and nodes per hidden layer). The values MSE and RMSE are used as selecting criteria for the optimal ANN configuration. Cross validations were used to find the optimum number of iterations (epochs) and MSE, RMSE were calculated again for every update of layer, node number as well as double-check of different weight function (Tables 2 and 3). The decreasing trend of the MSE and RMSE (tending to zero) suggested the suitability of the network to data (Fig. 4). Results emphasize the need to verify the accuracy of the selected model using internal and external validations prior to make any recommendation on the accuracy of the predictive model.

Tables 3a, b also showed the MSE and RMSE values obtained by 5 different weight functions. Depending on each chosen configuration, we can see that exponential function in general resulted in a higher MSE and RMSE when compared with other functions for data processing in different architectures. The linear function resulted in also high MSE and RMSE for all settings suggesting its non-suitability for a good model. Table 3a, b also indicated that the *Tanh sigmoid function* (hyperbolic tangent function) can be the best fitting function for the process as it led to the low values of MSE and RMSE in comparing with other functions for all developed architectures.

Our training steps led to the best configuration of ANN network is with 2 hidden layers (2 W and 1 F layers) as shown in Table 3a and 3b with the lowest MSE and RMSE (the yellowed parts). As the training revealed that the MSE and RMSE dropped as the epoch size was increased (Fig.4). However, to perform efficiently the simulation, this configuration should be operated at 45,000 epochs where MSE and RMSE reached the lowest values for both datasets. Table 4 shows all chosen parameters for our model configurations.

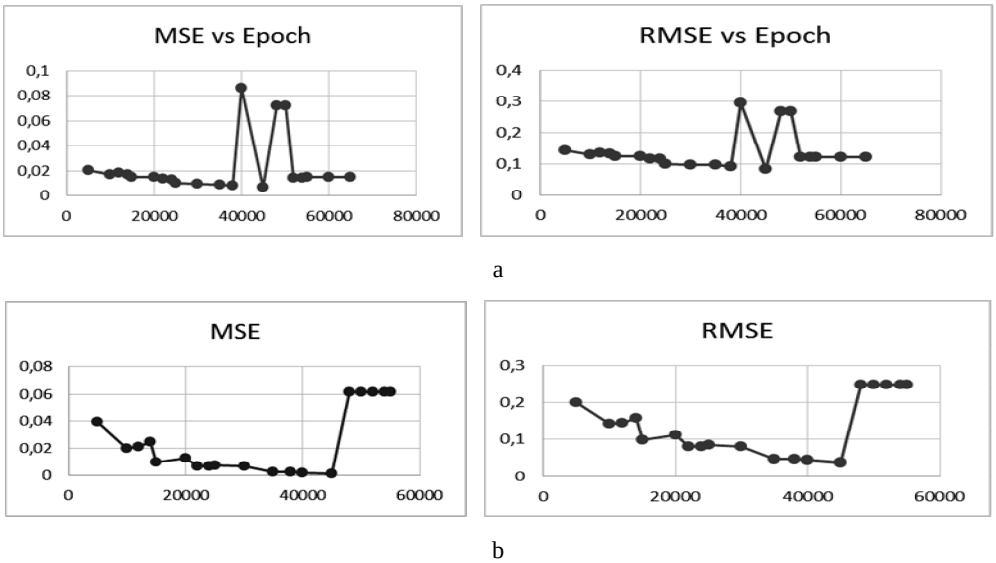


**Fig. 4.** Relationship between MSE and RMSE versus epoch (a) For McLaughlin data. (b) For Irishtown data

### 3.2. PC and Chl-a simulations versus input parameters

With the ANN structure adopted as the 'optimum' model after the training step, we started with the simulation for PC and Chl-a with different input parameters TP, Nitrates, pH, DO and temperatures. The series of figures from 5a, b to 9a, b showed our simulated PC and Chl-a versus these parameters. It is observed that in a general trend we obtained the good fitting results between experimental data and simulated values.

For the Chl-a values, the model developed the good agreement between simulated and real observation for both datasets of IR and McL. As shown in the left series of figures from 5a, b to 9a, b, we can see a good fit between Chl-a simulated and sampled data. It is also observed that the stronger Chl-a the better fitting shape we obtain for Chl-a values.

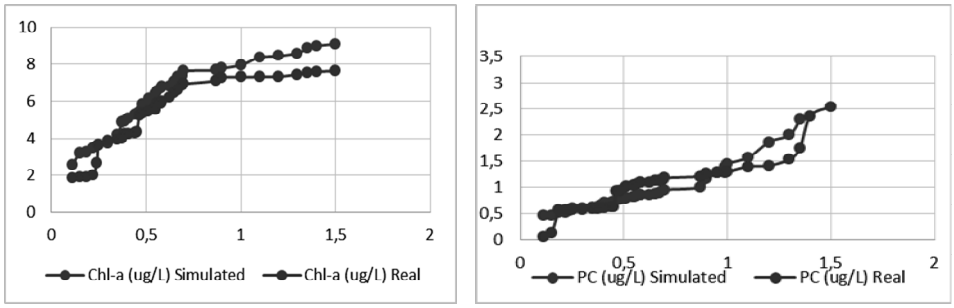


**Fig. 5a.** Measured and simulated data for Chl-a (left) and PC (right) vs Total Phosphorus for McL

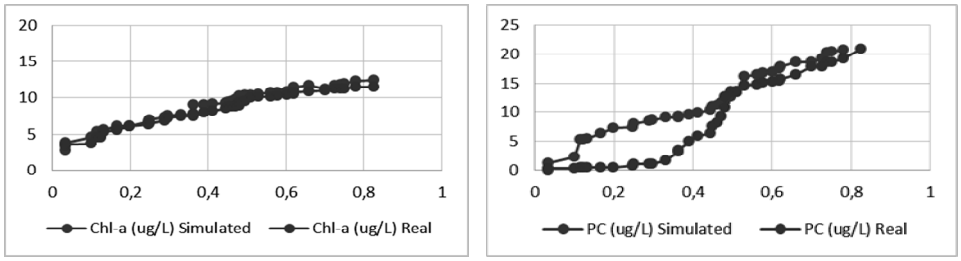


**Fig. 5b.** Measured and simulated data for Chl-a (left) and PC (right) *vs* Total Phosphorus for Irishtown

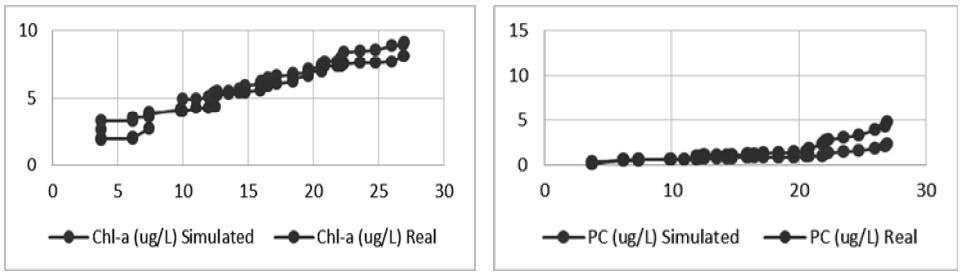


**Fig. 6a.** Measured and simulated data for Chl-a (left) and PC (right) vs Temperature for McL

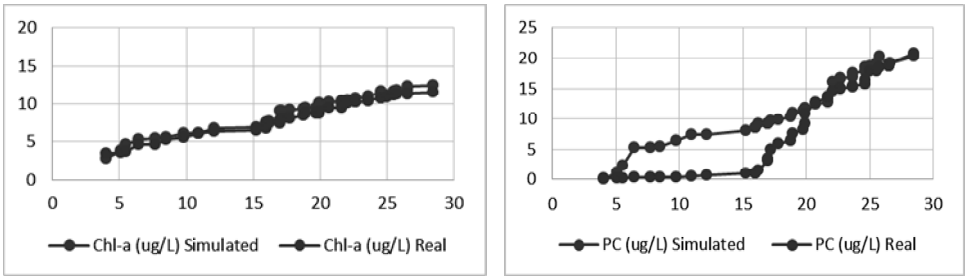


**Fig. 6b.** Measured and simulated data for Chl-a (left) and PC (right) vs Temperature for Irishtown

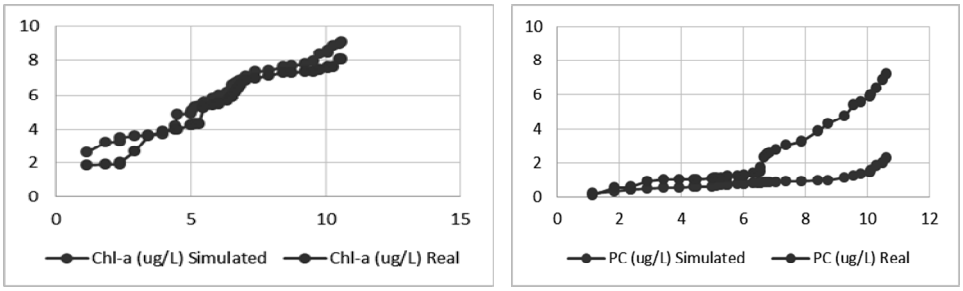


**Fig. 7a.** Measured and simulated data for Chl-a (left) and PC (right) *vs* DO for McLaughlin

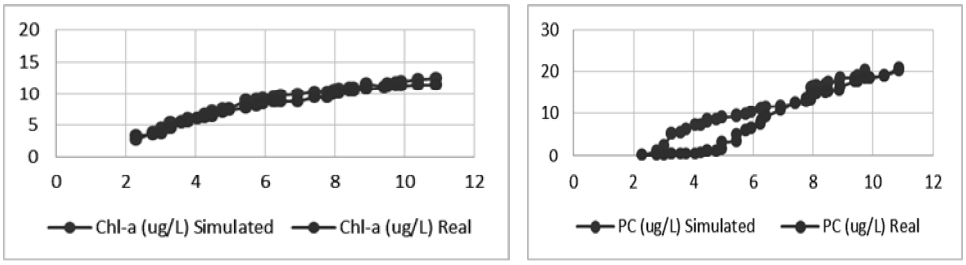


**Fig. 7b.** Measured and simulated data for Chl-a (left) and PC (right) vs DO for Irishtown

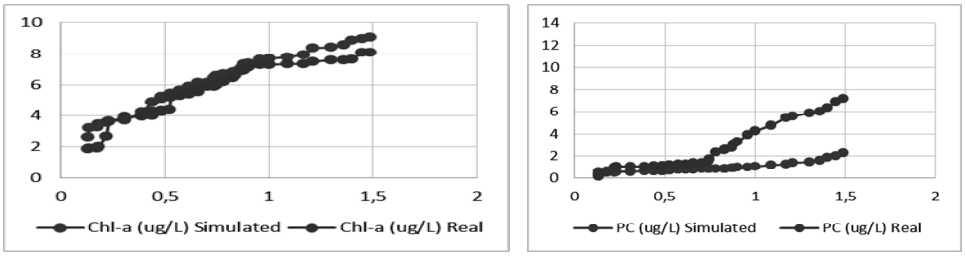


**Fig. 8a.** Measured and simulated data for Chl-a (left) and PC (right) vs Nitrates for McLaughlin

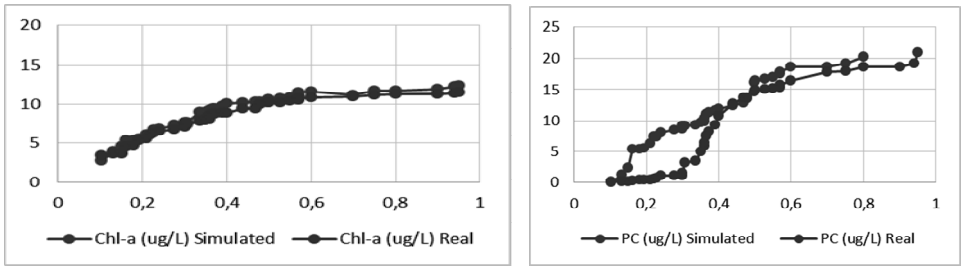


**Fig. 8b.** Measured and simulated data for Chl-a (left) and PC (right) *vs* Nitrates for Irishtown

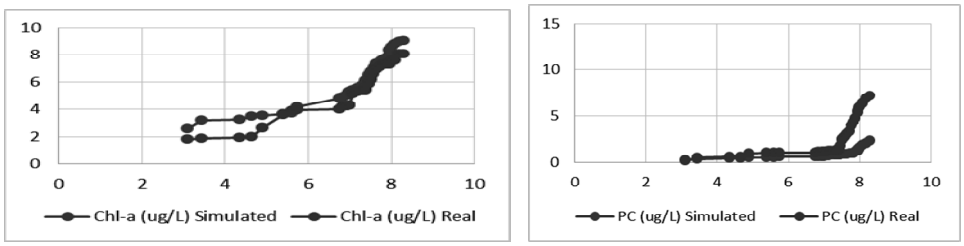


**Fig. 9a.** Measured and simulated data for Chl-a (left) and PC (right) vs pH for McLaughlin

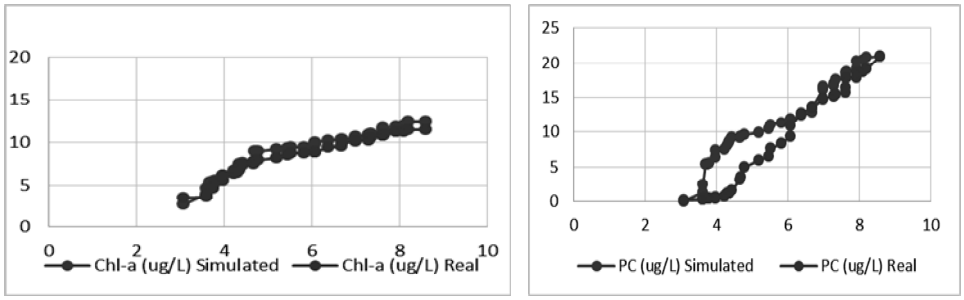


**Fig. 9b.** Measured and simulated data for Chl-a (left) and PC (right) vs *pH* for Irishtown

For the PC concentration, however, there exist some discrepancies between them. The simulated PC for Irishtown dataset makes a good agreement with the experimental data for all 5 input parameters (Fig. 5*b* to Fig. 9*b*, right side). But for McLaughlin dataset (Fig. 5*a* to Fig. 9*a*, right side), the agreement is not good as the difference between the simulated values and experimental ones, especially when PC versus Nitrates and DO (Fig. 7*a* right and Fig. 8*a* right) are very noticeable. This could be explained by following reasons:

- Data are not enough just with over a 2-year period amount of data
- Biological and physical reasons that we didn't include yet in our model.

- Errors from sampling and analysing
- To discuss about the performance of our ANN model suggested, we refer to Table 5 where the 3 indices R2, MSE and RMSE are shown. A lower MSE or RMSE values indicates that ANN model predictions is closely matching the real observation when compared to ANN model predicted values. Moreover, $R^2$ values are approaching to 1 which shows that ANN model predicted results are accurate.

To consider the combined effects of TP and Nitrates (Fig. 10 and 11), we can see that as the TP increases, so do the concentrations of Chl-a and PC. However, Chl-a and PC increase quite slowly with Nitrates (Fig. 10) for McLaughlin reservoir. In the case of Irishtown, Chl-a seemed not affected much by Nitrates (Fig. 11 above), PC was almost decreasing with Nitrates as shown in Fig. 11 (below).

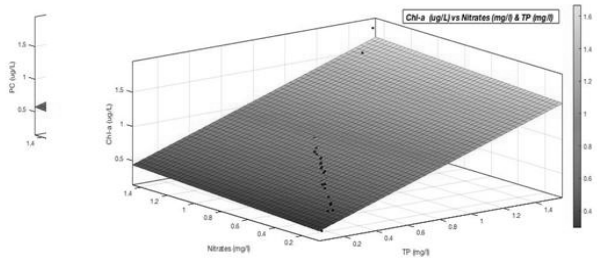


**Fig. 10:** The predicted results for PC vs Nitrates and TP ($R^2$ = = 0.9913 - above) and Chl-a vs Nitrates and TP ($R^2$ 0.9557- below) for McLaughlin reservoir

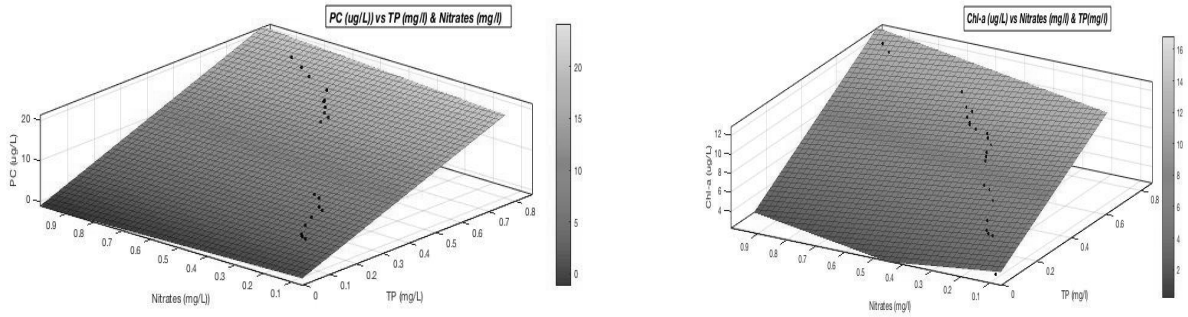


**Fig. 11.** The predicted results for PC vs Nitrates and TP ($R^2$ = 0.9607 – above) and Chl-a vs Nitrates and TP ($R^2$ = 0.9737 – below) for Irishtown reservoir

*Table 1a*

**Statistical Summary of Selected Data for the ANN modeling of McLaughlin reservoir**

| **McLaughlin data for training** | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Chl –a (ug/L) | PC(ug/L) | TP (mg/L) | DO(mg/L) | Temp (C) | pH | Nitrates (mg/L) |
| Min | 1.2 | 0 | 0 | 0.09 | 2.5 | 6.74 | 0 |
| Max | 10.53 | 11.73 | 0.693 | 10.5 | 27 | 8.04 | 0.88 |
| mean | 5.16 | 0.74 | 0.14 | 7.54 | 20.25 | 7.24 | 0.095 |
| Sd | 2.33 | 1.94 | 0.160616755 | 2.19 | 5.48 | 0.213 | 0.14 |
| Cv | 45.23 | 260.84 | 112.10 | 29.11 | 27.09 | 2.95 | 146.32 |
| Skewness* | 0.449 | 4.85 | 1.63 | -1.32 | -0.98 | 0.48 | 2.76 |
| Count | 79 | 79 | 79 | 79 | 79 | 79 | 79 |
| **validation** | | | | | | | |
| Min | 0.76 | 0.06 | 0 | 7.61 | 9.4 | 6.76 | 0 |
| Max | 14.86 | 10.86 | 0.495 | 10.66 | 22.4 | 7.61 | 0.5 |
| mean | 4.72 | 3.18 | 0.132 | 9.78 | 13.8 | 7.15 | 0.062 |
| Sd | 2.88 | 3.242 | 0.154 | 0.63 | 3.97 | 0.24 | 0.089 |
| Cv | 61.210 | 101.83 | 116.45 | 6.4643 | 28.8 | 3.452 | 143.22 |
| Skewness* | 1.32 | 0.80 | 1.0325 | -1.55 | 1.40 | 0.105 | 3.82 |
| Count | 33 | 33 | 33 | 33 | 33 | 33 | 33 |

*Table 1b*

**Statistical Summary of Selected Data for the ANN modeling of Irishtown reservoir**

| **Irish town data for training** | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Chl-a (ug/L) | PC (ug/L) | TP(mg/L) | DO(mg/l) | Temp (C) | pH | Nitrates (mg/L) |
| Min | 3.41 | 0.17 | 0 | 0.51 | 9 | 2.45 | 0 |
| Max | 45.05 | 28.17 | 0.825 | 10.39 | 28.5 | 8.6 | 0.6 |
| Mean | 13.207 | 8.356 | 0.132 | 7.78 | 22.75 | 7.2492 | 0.061 |
| Sd | 6.248 | 8.95 | 0.215 | 2.86 | 4.62 | 1.06 | 0.099 |
| Cv | 47.31 | 107.14 | 162.92 | 36.85 | 20.3 | 14.68 | 161.60 |
| Skewness* | 2.61 | 0.598 | 1.905 | -1.37 | -1.67 | -3.43 | 3.73 |
| Count | 52 | 52 | 52 | 52 | 52 | 52 | 52 |
| **validation** | | | | | | | |
| Min | 3.61 | 0.61 | 0 | 0.4 | 10.2 | 7.1 | 0 |
| Max | 55.75 | 30.93 | 1.16 | 12.64 | 22.3 | 9.22 | 0.12 |
| mean | 12.89 | 6.837 | 0.182 | 9.084 | 13.45 | 7.48 | 0.05 |
| Sd | 11.82 | 7.36 | 0.2604 | 3.856 | 3.78 | 0.523 | 0.044 |
| Cv | 91.68 | 107.74 | 143.091 | 42.45 | 28.12 | 6.99 | 86.58 |
| Skewness* | 2.66 | 2.11 | 3.019 | -0.97 | 1.36 | 2.52 | -0.059 |
| Count | 21 | 21 | 21 | 21 | 21 | 21 | 21 |

** Skewness characterizes the degree of asymmetry of a distribution around its mean.*

*Table 2*

**Developed networks using Tanh Sigmoid function at different time steps (epoch size) from 5000**

| | **Irishtown** | | **McLaughlin** | |
|---|---|---|---|---|
| **Epoch** | **MSE** | **RMSE** | **MSE** | **RMSE** |
| 5000 | 0.039 | 0.198 | 0.020 | 0.144 |
| 10000 | 0.020 | 0.142 | 0.017 | 0.130 |
| 12000 | 0.021 | 0.145 | 0.018 | 0.135 |
| 14000 | 0.025 | 0.158 | 0.017 | 0.131 |
| 15000 | 0.009 | 0.098 | 0.015 | 0.123 |
| 20000 | 0.012 | 0.111 | 0.015 | 0.124 |
| 22000 | 0.006 | 0.080 | 0.013 | 0.116 |
| 24000 | 0.006 | 0.080 | 0.013 | 0.115 |
| 25000 | 0.007 | 0.084 | 0.010 | 0.100 |
| 30000 | 0.006 | 0.080 | 0.009 | 0.096 |
| 35000 | 0.002 | 0.047 | 0.009 | 0.096 |
| 38000 | 0.002 | 0.047 | 0.008 | 0.091 |
| 40000 | 0.002 | 0.044 | 0.086 | 0.293 |
| 45000 | 0.001 | 0.037 | 0.007 | 0.083 |
| 48000 | 0.061 | 0.247 | 0.071 | 0.268 |
| 50000 | 0.061 | 0.247 | 0.071 | 0.268 |
| 52000 | 0.061 | 0.247 | 0.014 | 0.121 |
| 54000 | 0.061 | 0.247 | 0.014 | 0.120 |
| 55000 | 0.061 | 0.247 | 0.014 | 0.122 |
| 60000 | ----- | ---- | 0.014 | 0.121 |
| 65000 | ----- | ---- | 0.014 | 0.1220 |

*Table 3a*

**Tested mathematical functions to process the normalized data for McL**

| | Tanh Sigmoid | | Exponential | | Linear | | Logistic Sigmoid | | Morlet | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Model Structure** | **MSE** | **RMSE** | **MSE** | **RMSE** | **MSE** | **RMSE** | **MSE** | **RMSE** | **MSE** | **RMSE** |
| 2 W (5/4 and 4/2) and 1 F (2/2)* | 0.007 | 0.084 | 0.283 | 0.532 | 0.072 | 0.268 | 0.283 | 0.532 | 1.290 | 1.136 |
| 2 W (5/6 And 6/2) and 2 F (6/6 and 2/2) | 0.015 | 0.121 | 0.283 | 0.532 | 0.072 | 0.268 | 0.106 | 0.326 | 1.296 | 1.138 |
| 2 W (5/6 And 6/2) and 2 F (6/6 and 2/2) | 0.015 | 0.122 | 0.012 | 0.110 | 0.072 | 0.268 | 0.017 | 0.132 | 0.089 | 0.298 |
| 2 W (5/8 and 8/2) and 2 F (8/8 and 2/2) | 0.015 | 0.124 | 1.820 | 1.349 | 0.311 | 0.558 | 0.732 | 0.856 | 0.762 | 0.873 |
| 2 W (5/8 and 8/2) and 2 F (8/8 and 2/2) | 0.015 | 0.124 | 0.613 | 0.783 | 0.370 | 0.608 | 0.594 | 0.771 | 1.130 | 1.063 |
| 2 W (5/10 and 10/2) and 2 F (10/10 and 2/2) | 0.010 | 0.101 | 0.720 | 0.849 | 0.235 | 0.485 | 0.617 | 0.785 | 1.190 | 1.091 |
| 2 W (5/10 and 10/2) and 2 F (10/10 and 2/2) | 0.009 | 0.097 | 0.736 | 0.858 | 0.415 | 0.644 | 0.649 | 0.806 | 0.861 | 0.928 |
| 2 W (5/10 and 10/2) and 2 F (10/10 and 2/2) | 0.009 | 0.096 | 0.515 | 0.718 | 0.275 | 0.524 | 0.714 | 0.845 | 0.959 | 0.979 |
| 2 W (5/10 and 10/2) and 2 F (10/10 and 2/2) | 0.086 | 0.293 | 0.643 | 0.802 | 0.203 | 0.451 | 0.710 | 0.843 | 0.982 | 0.991 |

**Chosen structure*

*Table 3b*

**Tested mathematical functions to process the normalized data for Irishtown**

| | Tanh Sigmoid | | Exponential | | Linear | | Logistic Sigmoid | | Morlet | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Model Structure** | **MSE** | **RMSE** | **MSE** | **RMSE** | **MSE** | **RMSE** | **MSE** | **RMSE** | **MSE** | **RMSE** |
| 2 W (5/4 and 4/2) and 1 F(2/2)* | 0.0014 | 0.038 | 0.015 | 0.122 | 0.062 | 0.248 | 0.017 | 0.130 | 0.127 | 0.356 |
| 2 W (5/6 And 6/2) and 2 F(6/6 and 2/2) | 0.062 | 0.248 | 0.019 | 0.137 | 0.062 | 0.248 | 0.019 | 0.138 | 0.121 | 0.348 |
| 2 W (5/6 And 6/2) and 2 F(6/6 and 2/2) | 0.019 | 0.138 | 0.024 | 0.155 | 0.062 | 0.248 | 0.008 | 0.089 | 0.122 | 0.349 |
| 2 W (5/8 and 8/2) and 2 F(8/8 and 2/2) | 0.010 | 0.099 | 0.005 | 0.069 | 0.062 | 0.248 | 0.021 | 0.146 | 0.118 | 0.344 |
| 2 W (5/8 and 8/2) and 2 F(8/8 and 2/2) | 0.013 | 0.112 | 0.004 | 0.065 | 0.062 | 0.248 | 0.023 | 0.152 | 0.123 | 0.351 |
| 2 W (5/10 and 10/2) and 2 F(10/10 and 2/2) | 0.007 | 0.085 | 0.208 | 0.456 | 0.062 | 0.248 | 0.017 | 0.129 | 0.122 | 0.349 |
| 2 W (5/10 and 10/2) and 2 F(10/10 and 2/2) | 0.007 | 0.081 | 0.035 | 0.188 | 0.062 | 0.248 | 0.020 | 0.141 | 0.127 | 0.356 |
| 2 W (5/10 and 10/2) and 2 F(10/10 and 2/2) | 0.002 | 0.047 | 0.053 | 0.230 | 0.062 | 0.248 | 0.016 | 0.124 | 0.127 | 0.356 |
| 2 W (5/10 and 10/2) and 2 F(10/10 and 2/2) | 0.002 | 0.045 | 0.007 | 0.082 | 0.062 | 0.248 | 0.008 | 0.089 | 0.127 | 0.356 |

**Chosen structure*

*Table 4*

**Proposed ANN model parameter settings**

| **McLaughlin** | |
|---|---|
| **Parameters** | **Settings** |
| Percentage of data for training step | 80 % |
| Optimum Epoch | 52000 |
| Percentage of data for verification step (Internal validation) | 20 % |
| Number of hidden layers | 2 |
| Nodes per hidden layer | 9 |
| Learning rate | 0.10 |
| Momentum | 0.70 |
| Mathematical Function | Tanh Sigmoid |
| External validation | Independent data set |

*Continuation of the table 4*

| **Irishtown** | |
|---|---|
| **Parameters** | **Settings** |
| Percentage of data for training step | 50% |
| Optimum Epoch | 48000 |
| Percentage of data for verification step (Internal validation) | 50% |
| Number of hidden layers | 2 |
| Nodes per hidden layer | 9 |
| Learning rate | 0.10 |
| Momentum | 0.70 |
| Mathematical Function | Tanh Sigmoid |
| External validation | Independent data set |

*Table 5*

**Prediction performance of our ANN Model**

| Dataset | Model Structure | $R^2$ | MSE | RMSE |
|---|---|---|---|---|
| **McLaughlin Reservoir** | | | | |
| Training Dataset | 2 W (5/4 and 4/2) and 1 F (2/2) layers | 0.654 | 0.007 | 0.084 |
| Internal Validation Dataset | | 0.307 | 0.00974 | 0.0986 |
| **Irishtown Reservoir** | | | | |
| Training Dataset | 2 W (5/4 and 4/2) and 1 F (2/2) layers | 0.545 | 0.00141 | 0.038 |
| Internal Validation Dataset | | 0.487 | 0.00833 | 0.0912 |

## Conclusion

The simulation results showed that ANN model is quite suitable for capturing the non-linearity of the relationship between variables and helping to determine a predictive solution for algal bloom. Although there exist some discrepancies between the field measured data and simulated results due to some above-mentioned reasons, we believe with more good and correct data, our model can be improved and go in a more precise and accurate way. Results also suggested that the ANN model can be effectively used for a predictive scenario of algal bloom occurrence via the predicted values of Chl-a and PC in terms of various input parameters, especially when we consider more environmental factors such as wind factors and light intensity in the future.

It is concluded that mathematical models therefore have many advantages for studying coupled effects with more realistic parameters involving in algal blooms from the environment: nutrients, light and temperature, as well as intrinsic factors such as metabolism and grazing, etc. Acquiring sufficient data is an extremely important key factor in both developing simulation models as well as making accurate predictions. Microalgae dynamics modelling is considered an effective tool for complementing the limitations of field and laboratory experiments and is an approach that can be used at a minimal cost. However, we need to continue more sampling next several years for a complete database of these reservoirs as well as to build other independent databases in other lakes for the good external validations of the model, and for a more advanced model of the prediction and prevention of Harmful Algal Bloom problem.

## Acknowledgements

TNQ acknowledges the Natural Science and Engineering Research Council of Canada via Discovery Grant NSERC RGPIN 03796, Canada Foundation for Innovation as well as Nova Scotia Research and Innovation Trust via Grant No 31188 for equipment. We acknowledge Heather Fraser, Water Management Coordinator of the City of Moncton her active collaboration and accompaniment during the field trips. Kateryna Huschchyna and Kayla McLellan from BBML are deeply acknowledged by their sample analysis and treatments.

## References


[1] Huang, W., & Foo, S. (2002). Neural network modeling of salinity variation in Apalachicola River. *Water Research*, *36*(1), 356–362.

[2] Maier, H. R., and Dandy, G. C. (2000). Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental modelling & software*, *15*(1), 101–124.

[3] McCulloch, W. S. and Pitts, W. (1943), "A logical calculus of the ideas immanent in nervous activity", *The bulletin of mathematical biophysics*, Vol. *5*, No. 4, pp. 115–133.

[4] Torrecilla, J. S., Otero, L., & Sanz, P. D. (2004). A neural network approach for thermal/pressure food processing. *Journal of Food Engineering*, *62*(1), 89–95.

[5] Madic, M. J., & Radovanović, M. R. (2011). Optimal selection of ANN training and architectural parameters using Taguchi method: A case study. FME Transactions, 39(2), 79–86.

[6] Elangasinghe, M. A., Singhal, N., Dirks, K. N., & Salmond, J. A. (2014). Development of an ANN–based air pollution forecasting system with explicit knowledge through sensitivity analysis. Atmospheric pollution research, 5(4), 696–708.

[7] Pandey, D. S., Das, S., Pan, I., Leahy, J. J., & Kwapinski, W. (2016). Artificial neural network based modelling approach for municipal solid waste gasification in a fluidized bed reactor. Waste management, 58, 202–213.

[8] ASCE Task Committee on Application of Artificial Neural Networks in Hydrology. (2000). Artificial neural networks in hydrology. II: Hydrologic applications. Journal of Hydrologic Engineering, 5(2), 124–137.

[9] Khademi, F., Akbari, M., Jamal, S. M., & Nikoo, M. (2017). Multiple linear regression, artificial neural network, and fuzzy logic prediction of 28 days compressive strength of concrete. Frontiers of Structural and Civil Engineering, 11(1), 90–99.