Vietnam Journal of Mechanics, Vol. 46, No. 1 (2024), pp. 80 – 92
DOI: https://doi.org/10.15625/0866-7136/20350

APPLICATION OF ANN FOR COASTAL WATER QUALITY PREDICTION: QUANG BINH CASE STUDY

Van Manh Dinh[1,2], **Chinh Kien Nguyen**[1,*], **Thi Huong Le**[1], **Thi Minh Hanh Pham**[1], **Thi Hang Nguyen**[1]
[1]*Institute of Mechanics, VAST, Hanoi, Vietnam*
[2]*VNU University of Engineering and Technology, Hanoi, Vietnam*
[*]E-mail: nckien@imech.vast.vn

Received: 11 January 2024 / Revised: 23 February 2024 / Accepted: 29 February 2024
Published online: 31 March 2024

Abstract. The coastal areas of Quang Binh province play a crucial role not only in the economic and tourism development of the coastal province but also in the overall development of the Northern Central region. Therefore, monitoring and forecasting seawater quality in this region is vital. However, the water quality condition assessment faces many limitations due to the lack of measured data and the complication of numerical models. Meanwhile, the artificial intelligence model for simulating and predicting water quality has been widely applied due to its timely and reliable calculating abilities. This research has piloted the prediction of some water pollution parameters on the coast of Quang Binh province using an artificial neural network (ANN) model. This presents a novel approach to identifying implicit relationships between variables based on data analysis techniques via ANN. An ANN model was built to analyze the measured environmental time series data at Dong Hoi station, Quang Binh province, from 2002 to 2022. The calculation results of the training (70% of the data set, NSE: 0.81) and testing (the rest 30% of the data set, NSE: 0.5) of the model have satisfied the total coliform parameter, indicating the promise of applying the ANN model for water quality prediction.

Keywords: prediction, pollution, total coliform, Quang Binh province.

1. INTRODUCTION

Water pollution consistently remains an urgent problem and receives a lot of attention because of its serious impacts on health as well as on the environment and ecosystem. Currently, the application of models to simulate water quality has become popular in many countries around the world. Water quality simulation models are considered

one of the most effective support tools for assessing and predicting water quality as well as assessing the spread and diffusion of pollutants during water pollution incidents. A number of water quality calculation models have been developed, such as QUAL, WASP, QUASAR, and MIKE Ecolab, with the advantage of being able to provide information to assess water quality spatially and temporally. However, setting up the model is relatively complicated, as it requires a substantial amount of measured data and consumes considerable time. The users also have to possess specialized knowledge in calibrating, testing, and evaluating model results.

Recently, ANN models have been widely applied in water quality prediction. Given a sufficient database and the advancement of computing technology, this type of model's ability is enable to provide highly accurate information with significant efficiency and in some cases can replace numerical hydraulics-environment models.

In practice, the ANN model possesses outstanding advantages and is suitable for effectively managing, evaluating, simulating, and predicting water quality in a number of countries around the world. However, utilizing artificial intelligence technology to simulate water quality is still one of the relatively new research approaches across the globe. A number of studies have been conducted in relation to the application of ANN models to forecast water quality variables in rivers and estuaries. Dogan et al., (2009) [1] succeeded in using ANN technique for modeling biological oxygen demand (BOD) in Melen River, Turkey; The ANN were used to develop models for prediction dissolved oxygen (DO) and specific conductance (SC) in Delaware river, United State (Heydari et al., 2013) [2]; The research results by Musavi and Golabi [3] revealed that ANNs can be used with more than 90% accuracy for simulating water variables such as CO_3, HCO_3, SO_4, Cl, Na, Ca, Mg, K, EC, TDS and SAR in Karoon river, Iran; Similarity, Singh et al. (2009) [4] concluded that the ANN models can be used as tools for the computation of DO and BOD concentrations in the Gomti river, India; Faruk [5] suggested that ANN can be adequate in modeling and predicting time series data of Boron (Bor), DO and water temperature in Büyük Menderes river, southwest Turkey, The research of Rajiv et al. [6] demonstrated that ANN not only can be used for predicting a single water quality variable but also for an integrated water quality index which based on several parameters: DO, pH, turbidity, *E. coli* and conductivity, ... ANN models are also used to quickly evaluate and forecast coastal water quality parameters such as salinity, temperature, dissolved oxygen and chlorophyll, ... as researched by Palani et al. [7, 8] in the mouth of the East Johor Strait, Singapore.

In Vietnam, up to now, there have not been many studies evaluating water quality simulation using ANN. One of the reasons is the lack of data for training. Among related studies, it is worth mentioning the research and development of a scientific basis for calculating surface water quality index using machine learning methods for some river

basins by An [9], and Hoai et al. [10]. Studies on building a machine learning model to calculate a number of water quality indicators with 4 main parameters: pH, BOD_5, PO_4^{3-}, coliform for the peninsula and coastal areas by Thai [11], Phong and Duong [12].

In this study, the authors will construct a water quality simulation tool based on an artificial neural network model. Accordingly, the research will use this technology to build a correlation between input (in situ or automatic measurement of oceanography and coastal marine environment variables) and output (laboratory analyzing marine water quality parameters) at the monitoring station in the coastal area of Quang Binh province. This relationship, after being established, will be used to simulate water quality, thus contributing to forecast and future pollution control. In comparison to traditional numerical models, artificial intelligence models do not require users to have in-depth professional knowledge because the analysis and data processing tasks are hidden behind the scenes. The ANN also provides quick calculation results, very suitable for water quality simulation and prediction as well as decision making support in pollution control activities.

The coastal waters of Quang Binh province are facing the danger of pollution and water quality degradation. One of the main causes is due to massive exploitation and aquaculture activities without comprehensive planning, leading to estuaries and coastal areas having to receive wastewater from different sources. The research will contribute to providing scientific, effective, and low-cost methods for calculating surface water quality indicators that suit the actual conditions of coastal areas of Quang Binh province.

2. METHODS AND DATA

2.1. Artificial neural network

Artificial Neural Network (ANN) is built from basic components that are artificial neurons with many inputs and one output Fig. 1. Each neuron simulates a biological neuron, including an activation threshold (bias) and an activation function (or transfer function), characterizing the properties of the neuron [13].

In which:

- x_i $(i = 1, m)$: The set of input signals of the neuron;

- w_{jk}: Set of links, each link between the j^{th} input signal and k neuron is represented by a weight, randomly initialized at the time of network initialization and continuously updated during the learning process;

- Σ: The totalizer is used to calculate the sum of the product of the inputs with their associated weights;

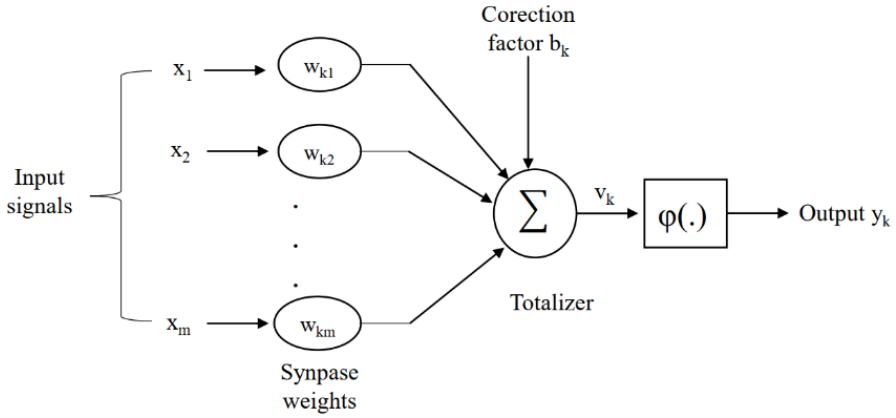

Fig. 1. The structure of an artificial neuron

- b_k: Bias, included as a component of the transfer function;

- $\varphi(.)$: Transfer function is also called Activation function. Input data of this function are the result of the totalizer and the given bias;

- y_k: The output signal of a neuron, with each neuron having at most one output.

Although each individual neuron can perform certain information processing functions, the power of neural computing is largely achieved by combining neurons in a unified architecture. Fig. 2 [13] shows the most widely used multi-layer feedforward network in ANN models.

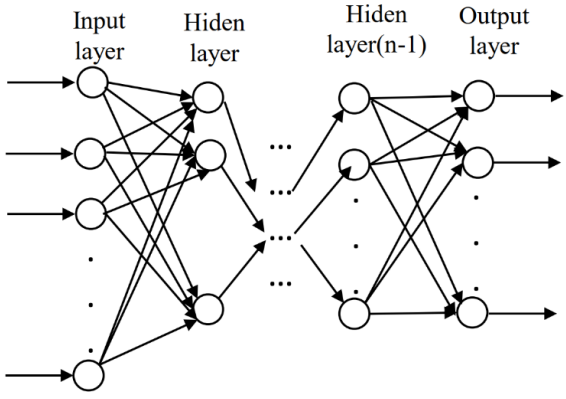

Fig. 2. Straight propagation neural network

The advantage of an artificial neural network is that it allows the construction of a computational model with a very high ability to learn from data. The neural network training process is based on the regression error between the calculated value and the

actual measured value. The training algorithm will adjust the connection weights of the neural network to minimize the regression error on the training samples. After the network is successfully trained, the weight matrix will be updated for use in the forecasting process.

2.2. Research area and data collection

Quang Binh is a coastal province located in the south of the Northern Central regions, Central of Vietnam (Fig. 3). Rivers and streams in the coastal districts of Quang Binh mostly originate in the province's territory and then flow directly into the East Sea. Due to the narrow and steep terrain rivers and streams here are often short, steep and the density of rivers and streams is quite high. The flow of rivers is relatively large. There are five main rivers: Ron River, Gianh River, Ly Hoa River, Dinh River, and Nhat Le River.

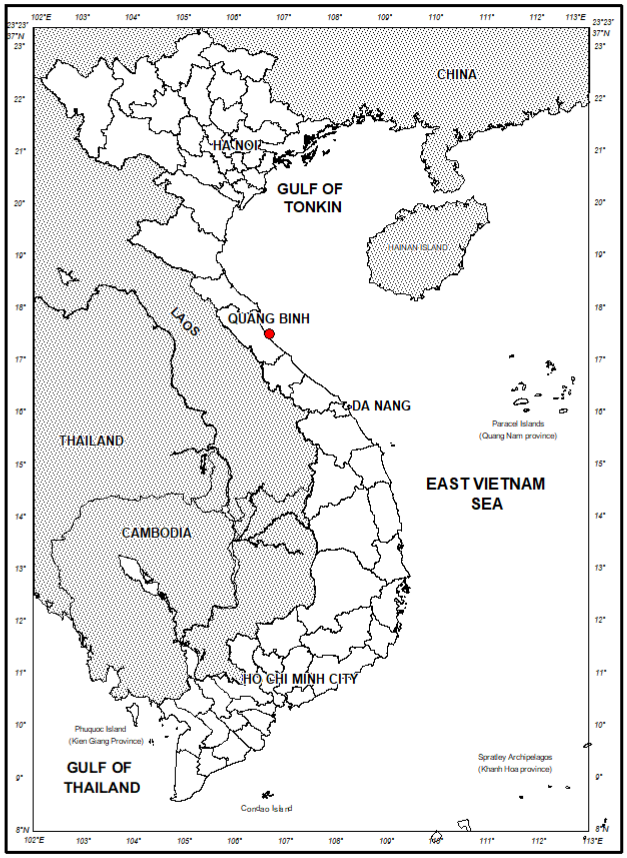

Fig. 3. Study area

Quang Binh is one of many beach tourism destinations in Vietnam that has been developing in recent years. However, the water resources of the coastal districts of Quang Binh province are affected by a number of activities with potential risks of marine environmental pollution such as maritime transportation, oil exploitation, and aquaculture, ... Besides, along the coast, there are small-scale exploitation and aquaculture activities, increasingly developing tourism activities and crowded residential areas.

In order to predict water quality for the coastal area of Quang Binh province by using an artificial neural network model, data was collected as input for the model from the results of the annual mission: "Marine environment monitoring in the Central part of Vietnam" hosted by the Institute of Mechanics, from 2002 to 2022 [14]. As follows:

- Study area: Dong Hoi station, coordinates 17°30'36"N, 107°31'30"E.

- Data: Input parameters as temperature, salinity, DO, pH, TSS, water level, and flow velocity; Output parameters as NO_3^-, PO_4^{3-}, Zn, oil and grease, total coliform in surface seawater, high tide and low tide phases. In which:

Surface seawater water temperature, salinity, DO and pH were measured in situ by the water quality probe WQC-22 A (TOA, Japan); Water level was specified by the tide harmonic constant method; Flow velocity was measured by the handheld 2-D electromagnetic current meter, model AEM 213-D (JFE Advantech, Japan).

Surface seawater samples (50 cm depth) were collected according to Circular No. 10/2021/TT-BTNMT - Environmental monitoring techniques and management of environmental quality monitoring information and data of Vietnam. Pre-treatment and preservation of samples was following the Standard methods for Examination of wastewater 15 Edition, 1990 APHA-AWWA-WPCF. Dissolved nutrients (NO_3^- and PO_4^{3-}) were analyzed spectrophotometrically using a UV-VIS spectrophotometer (DR6000, HACH, USA), following specific methods Ultraviolet Spectrophotometric Screening Method (APHA method 4500-NO_3-B) and the Ascorbic Acid Method (PO_4: 4500-P E), respectively. Each measurement was performed three times for the average reported value. Oil and grease were laboratory analyzed by TCVN 7875:2008 - Water - Determination of oil and grease - Partition-infrared method. Zn in the dissolved phase was analysed by an Agilent 8900 ICP-MS Triple Quadrupole, in MS/MS mode (Agilent Technologies, USA) following the SMEWW 3125:2012 method. Total suspended solid (TSS) concentration was determined by TCVN 6625:2000, ISO 11923:1997 - Water quality Determination of suspended solids by filtration through glass-fiber filters. Total coliform was determined by the TCVN 6187-1-1996 (ISO 9308-1-1990) Water quality - Detection and enumeration of coliform organisms thermotolerant coliform organisms and presumptive *Escherichia coli.* Part 1: Membrane filtration method.

Frequency: total of 51 monitoring campaigns (127 samples), including 4 monitoring campaigns/year in 2002-2004, 2006, 2020-2022; 3 campaigns/year in 2019; and 2 campaigns/year in 2005, 2007-2011, 2015-2018, more detail on monitoring campaigns and samples are as following (Table 1).

Table 1. Sampling time and number of samples at Dong Hoi monitoring station

Year	Sampling time and number of samples								Total samples
	Campaign No 1		Campaign No 2		Campaign No 3		Campaign No 4		
2002	19 Feb	2	13 May	2	9 Aug	2	30 Nov–1 Dec	4	10
2003	16 Mar	2	11 May	2	07 Aug	2	20 Nov	2	8
2004	27 Feb	2	21 May	2	18–19 Aug	4	15 Nov	2	10
2005	8–9 Mar	4			27–28 Aug	4			8
2006	26 Feb	2	24 May	2	28 Aug	2	9 Nov	2	8
2007	17–18 Mar	4			13 Aug	2			6
2008	8–9 Mar	4			13–14 Sep	4			8
2009	2–3 Apr	4			17–18 Aug	4			8
2010	20 May	2			25 Oct	2			4
2011	3–4 Apr	4			22–23 Aug	4			8
2015	17 Apr	2			20 Sep	2			4
2016	20 Apr	2			30 Sep	2			4
2017	9 May	2			27 Sep	2			4
2018	26 May	2			23 Aug	2			4
2019	6–8 May	3	10–12 Jul	3	28–30 Sep	3			9
2020	16 Mar	2	20 May	2	13 Aug	2	20 Nov	2	8
2021	15 Mar	2	24 May	2	30 Sep	2	17 Dec	2	8
2022	31 Mar	2	2 Jun	2	5 Aug	2	11 Nov	2	8

3. SETTING UP THE MODEL AND CALCULATION

3.1. Model setup

The steps to build an ANN model to predict the concentration of pollutants in coastal areas are described as follows:

Step 1: Determine the input parameters (temperature, salinity, DO, pH, TSS, water level, velocity) that are correlated with the output parameters (NO_3^-, PO_4^{3-}, Zn, mineral oil and grease, total coliform) that need to be predicted for the ANN model's calculation samples.

Step 2: Build the model:

Divide the sample statistics into two sets where most of the samples are used for calibration (called the "training set"), and the remaining samples are used for model validation and error estimation (called the "test set"). The fit of a model is related to the ratio of training samples to test samples, usually choosing a ratio of 2:1 [13]. Select input layer (temperature, salinity, DO, pH, TSS, water level, velocity), hidden layer, output (respectively $NO_3{}^-$, $PO_4{}^{3-}$, Zn, mineral oil and grease, total coliform), the activation functions used are Sigmoid functions.

Step 3: Calibrate the model:

To receive the optimal set of weights for the ANN model, the input data will be transmitted to the hidden layer through a combination of weights, each input will have a separate weight corresponding to each neuron on the hidden layer. After entering the hidden layer, the values will be calculated by algorithms to produce output results. These results are compared with pre-calculated values. If the error is large, the inverse transmission algorithm will be used to adjust the weights and continue to recalculate on the hidden layers until the error is minimized, giving the final output result.

Step 4: Test the model and evaluate:

After having the optimal set of weight matrices in step 3, perform calculations to "re-forecast" with the data set as the "test set". Representing the output values calculated by the model and real data measured in the study area.

The simulation effectiveness of the model is evaluated using statistical methods to compare the quality and reliability of simulation results with measured data. In this study, the statistical methods of the Nash-Sutcliffe efficiency index (NSE), standard deviation ratio index (RSR), and total balance coefficient (PBIAS) are used. The model is considered good and reliable when the NSE value is close to 1.0, RSR approaches 0, and PBIAS $\leq \pm 15\%$ [15].

3.2. Computational testing

In this study, the authors conducted research with 127 samples, 89 samples used to train and 38 others to test the model, with the set of weight matrices obtained after the training step, this is an approximately reasonable ratio according to the Pareto principle (with adjustment from the ratio 80/20 to 70/30 to suit the existing sample volume).

However, with the established artificial neural network model and the input of a standardized data set, the calculation results show that the model is too close to the data (overfitting phenomenon). This will be accurate on the training set but on the test set the results are unacceptable. This model usually has small errors and large dispersion. Each output parameter ($NO_3{}^-$, $PO_4{}^{3-}$, Zn, oil and grease mineral, total coliform) was tested

for thousands of calculations with initial sets of randomly derived weight matrices. The results all gave the NSE indexes greater than 0.8 in the training step. However, the NSE indexes of the testing step were all less than 0.5 (negative values in most cases). The reason is that, when training the model on a lot of noisy data the model becomes too complex compared to the necessary leading to failure in generalizing. So when encountering new data it will predict incorrectly. Several options have been chosen by the authors to overcome this phenomenon, among which is reducing the complexity of the neural structure (from 2 hidden layers to 1 hidden layer, reducing the number of nodes of a hidden layer from 9 down to 7) and combined with the "early stopping" technique (when training the model, the loss function of training set and test set does not always decrease at the same time. Up to a certain epoch, the loss function of the training set will continue to decrease, but the loss function of the test set will not decrease but increase again. So to prevent it - the overfitting phenomenon, model training will stop at that moment).

Eliminate input data such as DO and pH in the measured data chain; Eliminate error or too different samples (0 values or some values are too far from the remaining measured value range) for each indicator that needs to be predicted: $NO_3{}^-$, $PO_4{}^{3-}$, Zn, oil and grease minerals, total coliform. However, only the model, that calculates total coliform, is effective, other substances still cannot escape the overfitting state.

3.2.1. *Training results*

To evaluate the calculation results, the author uses the NSE efficiency index, RSR correlation index, and PBIAS to compare the measured total coliform and total coliform of the model training (Fig. 4). The results show that the NSE index is good (> 0.75), the RSR index is low as well as the PBIAS index $< 20\%$, showing that the model after

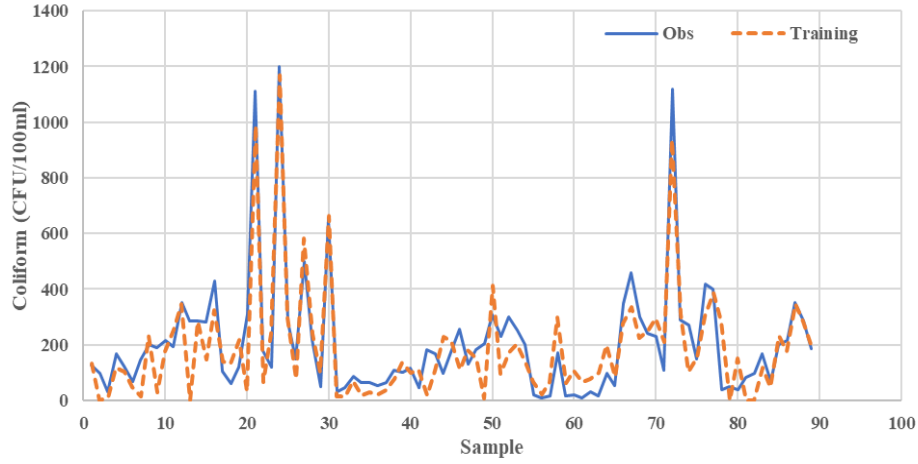

Fig. 4. Graph comparing total coliform values of model training and measured data

training is highly reliable. In addition, the graph also shows that the maximum and minimum values are also simulated by the model quite well compared to the measured data (Table 2).

Table 2. Results of model evaluation indicators for the training plan

Index	NSE	RSR	PBIAS (%)
Training	0.81	0.43	9.46

3.2.2. *Test results*

After finding the best possible set of weight matrices in the training step, the authors used this set of matrices with the test sample data set to test the model. Fig. 5 is the result of comparing the total coliform value obtained from the testing process of the ANN model with the measured data series.

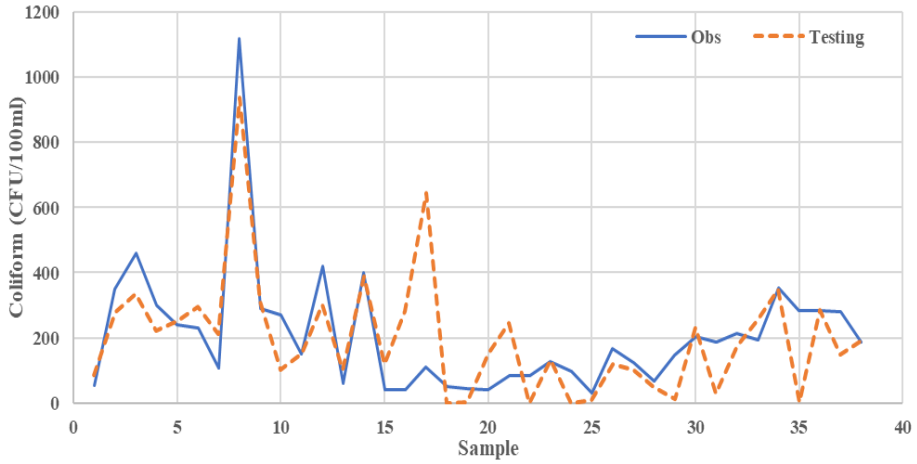

Fig. 5. Graph comparing the total coliform value of the model test and measured data

The values of the results evaluation indexes are all at the pass level, NSE = 0.5 is at the acceptable level, and the RSR or PBIAS indexes also give good results (Table 3). However, at some points, there is a certain difference between the calculated and measured values, possibly due to too little data to train the model as well as the measured data of the test series encountering certain noise. But in general, the graph shows a certain consistency in value between the two data series: calculated from the tested ANN model and measured value.

Table 3. Results of model evaluation indicators for the test plan

Index	NSE	RSR	PBIAS (%)
Test	0.5	0.71	5.23

4. CONCLUSIONS

In this study, measured data at Dong Hoi station - Quang Binh from 2002 to 2022 was used as input for an artificial neural network model to predict the values of main environmental variables. The contributions of this study can be summarized by the following points:

- Successfully built an artificial neural network and trained to calculate total coliform density in coastal seawater with good results. The optimal matrix obtained after the training process is used to predict the environment's total coliform with quite good results.

- The established model was not able to calculate some other pollutants such as $NO_3{}^-$, $PO_{43}{}^-$, Zn, mineral oil, and grease because of the overfitting phenomenon.

The fact that the ANN model can predict the total coliform index quite well but not well for other indices such as $NO_3{}^-$, $PO_{43}{}^-$, Zn, mineral oil, and grease can be explained as follows: when setting up the ANN model in the study area, the input data of the new model only includes a few parameters such as temperature, salinity, DO, pH, TSS, water level and flow velocity. While coliform density is quite sensitive, as it depends heavily on environmental factors such as temperature, salinity, DO, and pH; nutrients (nitrogen, phosphorus) and heavy metals are less affected by these factors. These parameters depend directly on supplies from the mainland, estuarine areas, and even bottom sediments. Therefore, to improve simulation efficiency and model reliability for total coliform as well as other environmental factors, it is necessary to research and selectively choose additional input data for the model.

Regarding the overfitting phenomenon, the future research direction is to apply a number of solutions to fix such as: collecting more data or creating enhanced data (for example, converting flow velocity in velocity and angle values); Using innovative techniques such as regularization and dropout techniques in artificial neural network algorithms; Identify the main factors affecting the output qualities (using the game theory method SHAP - SHApley Additive exPlanations to interpret the output of the machine learning model).

The initial results of the study show that the application of the ANN model in coastal water quality prediction is a potential new direction. However, it is still necessary to

improve the ANN model as well as collect more measured data so that the model can better reflect the relationship between input and output quantities. With the advantages of ANN models, it will help to have more tools to support managers in operating in the fields of tourism industry, fisheries as well as developing economic sectors and protecting the environment.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ACKNOWLEDGMENT

The authors would like to thank the financial support of the Vietnam Academy of Science and Technology (VAST) for the Project "Detailed investigation of topography, hydrodynamics and marine environment in Quang Binh", No. CT0000.04/21-22.

REFERENCES

[1] E. Dogan, B. Sengorur, and R. Koklu. Modeling biological oxygen demand of the Melen River in Turkey using an artificial neural network technique. *Journal of Environmental Management*, **90**, (2009), pp. 1229–1235. https://doi.org/10.1016/j.jenvman.2008.06.004.

[2] M. Heydari, E. Olyaie, H. Mohebzadeh, and Ö. Kisi. Development of a neural network technique for prediction of water quality parameters in the Delaware River, Pennsylvania. *Middle-East Journal of Scientific Research*, **13**, (2013), pp. 1367–1376.

[3] S. H. Musavi-Jah and M. Golabi. Application of artificial neural networks in the river water quality modeling: Karoon River, Iran. *Journal of Applied Sciences*, **8**, (2008), pp. 2324–2328. https://doi.org/10.3923/jas.2008.2324.2328.

[4] K. P. Singh, A. Basant, A. Malik, and G. Jain. Artificial neural network modeling of the river water quality—A case study. *Ecological Modelling*, **220**, (2009), pp. 888–895. https://doi.org/10.1016/j.ecolmodel.2009.01.004.

[5] D. Ömer Faruk. A hybrid neural network and ARIMA model for water quality time series prediction. *Engineering Applications of Artificial Intelligence*, **23**, (2010), pp. 586–594. https://doi.org/10.1016/j.engappai.2009.09.015.

[6] R. Gupta, A. N. Singh, and A. Singhal. Application of ANN for water quality index. *International Journal of Machine Learning and Computing*, **9**, (2019), pp. 688–693. https://doi.org/10.18178/ijmlc.2019.9.5.859.

[7] P. Sundarambal, S. Y. Liong, P. Tkalich, and P. Jegathambal. Development of a neural network model for dissolved oxygen in seawater. *Indian Journal of Marine Sciences*, **38**, (2), (2009), pp. 151–159.

[8] S. Palani, S.-Y. Liong, and P. Tkalich. An ANN application for water quality forecasting. *Marine Pollution Bulletin*, **56**, (2008), pp. 1586–1597. https://doi.org/10.1016/j.marpolbul.2008.05.021.

[9] H. T. An. Research combines hydraulic and artificial intelligence models to simulate the water quality of the Nhu Shi-Dai River. *Journal of Hydrological Meteorology*, **739**, (2022), pp. 67–80.

[10] P. N. Hoai, P. B. Quoc, and T. Thanh Thai. Apply machine learning to predict saltwater intrusion in the Ham Luong river, Ben Tre province. *VNU Journal of Science: Earth and Environmental Sciences*, **38**, (2022). https://doi.org/10.25073/2588-1094/vnuees.4852.

[11] T. H. Tran. Application of MIKE 21 FM modelling to simulate the water quality at the coastal area Đình Vũ. *Science and Technology Development Journal - Natural Sciences*, **1**, (2017), pp. 282–292. https://doi.org/10.32508/stdjns.v1it4.470.

[12] N. D. Phong and H. H. Duong. Applied studies of machine learning models to predict surface water quality indicators in the Ca Mau Peninsula. *Journal of Hydraulic Science and Technology*, **76**, (2023), pp. 1–12.

[13] C. K. Nguyen, T. H. D. Thi, H. N. Thi, T. H. Do, and T. D. Nguyen. Applying gridded rainfall data to calculate inflow discharge into Ban Chat hydropower reservoir basin. In *Proceedings of the 7th International Conference on Engineering Mechanics and Automation (ICEMA7)*, Publishing House for Science and Technology, ICEMA 2023, (2023), pp. 147–152. https://doi.org/10.15625/vap.2023.0136.

[14] Institute of Mechanics. *Annual Summary Report "Marine environment monitoring in the Central part of Vietnam" from 2002 to 2022*.

[15] D. N. Moriasi, J. G. Arnold, M. W. V. Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, **50**, (3), (2007), pp. 885–900. https://doi.org/10.13031/2013.23153.