Analysis of long-term performance of full-scale reverse osmosis desalination plant using artificial neural network and tree model

Yongjun Choi[1], Younggeun Lee[2], Kwanghee Shin[2], Youngkyu Park[1], Sangho Lee[1†]

[1]School of Civil and Environmental Engineering, Kookmin University, Jeongneung-Dong, Seoul, 02707, Korea
[2]Corporate R&D Institute, Doosan Heavy Industries & Construction, Co., Ltd., 16858, Korea

ABSTRACT

The reverse osmosis (RO) technology is currently the leading desalination method. However, until recently, application of RO technology on a large scale has been primarily limited by membrane fouling. The mechanism of fouling is complex, which is not well understood in full-scale plants. Although many studies about modeling and prediction of fouling have been done, in most cases, the experimental data set of lab or pilot scale systems, which may not show fouling characteristics well in full-scale systems were used. In this study, both artificial neural network (ANN) model and tree model (TM) was evaluated to analyze long-term performance of full scale reverse osmosis desalination plant. The results of application of the ANN and TM indicated high correlation coefficients between the measured and simulated output variables. However, it is not easy to use ANN for the full scale plant operation because the final model is not expressed as a form of mathematical functions. TM has advantages over ANN because the model can be obtained as forms of simple function and it showed reasonably high R^2. Therefore, TM is shown to be more adequate than ANN for developing models in which the full-scale RO plant data is considered as an input.

Keywords: Artificial neural network, Fouling, Reverse osmosis, Statistical analysis, Tree model

1. Introduction

Water scarcity is regarded as the largest global risk facing mankind [1]. The World Health Organization predicts that by 2050, four billion of us nearly two thirds of the world's present population will face severe fresh water shortages [2]. Producing drinking water from seawater has been technologically achievable for several decades in many countries [3]. Today there are two primary technologies used for desalination, thermal distillation (multi-stage flash distillation, multi-effect distillation) and membrane separation (reverse osmosis (RO), nanofiltration (NF)). Among them, RO technology has now evolved into the leading desalination technology globally [4, 5].

However, membrane fouling is still one major hurdle to application of RO system [5]. Membrane lifetime, feed pressure and salt rejection are primarily affected by the fouling at the membrane surface [6, 7]. Once membrane fouling occurs, it will cause a higher energy use, a higher cleaning frequency and a shorter life span of the membrane [8]. The impact of membrane fouling on the RO operation is more serious in large-scale desalination plants because it leads to an increased cost for their operation and maintenance. Accordingly, it is necessary

to analyze and predict membrane fouling in large-scale RO desalination plants. Nevertheless, modeling and prediction of membrane fouling is a major challenge in the long-term operation of the full-scale plants due to its complexities. Although many studies about modeling and prediction of fouling have been done, in most cases, the experimental data set of lab or pilot scale systems, which may not show fouling characteristics well in full-scale systems were used [6-11].

Modeling of membrane fouling in a large-scale RO process is quite different from those in a lab or pilot scale systems. In fact, a variety of theoretical models that employ hydrodynamic, chemical, and physical inputs in physical equations have been proposed to predict membrane fouling in RO process [9-11]. Since membrane fouling is an extremely complex phenomenon that has not been defined precisely, these models simplify the fouling phenomenon and could only predict the fouling behavior induced by feed water containing relatively simple foulants, such as mono-disperses, colloids, calcium sulfate or calcium phosphate. These approaches have been found to be successful in laboratory and pilot scales. However, they are relatively unsuitable for predicting membrane performance in full-scale plants [9-13].

Recently, machine learning (ML) techniques, such as artificial

neural network (ANN), genetic algorithm optimization, genetic programming and tree model (TM) have been increasingly used in membrane systems due to their ability to model and analyze complex problems that were previously difficult or impossible to solve [14]. Among many ML techniques, ANN is the most widely used to prediction of fouling [9, 14, 15]. ANN is numeric technique able to capture complex input-output relationships because of their ability to learn linear, as well as nonlinear correlative patterns between set of input data and corresponding target values, directly from available experimental data [16].

However, one of the disadvantages of ANN is that for a decision maker it is very difficult to analyze the structure of the resulting ANN and to relate it to the outputs [15]. But the relation rules between input and output can be explicitly observed from the TM. TM is a generalization of decision tree which is widely used in solving classification problems and more specifically very common in data mining applications. Whereas decision tree only handles qualitative or discrete-value attributes, TM can deal with continuous values. TM is a data-driven algorithm built of a rule based predictive structure using a top-down induction approach [16].

The objective of this study is to investigate the comparative performance of ANN and TM to analyze long-term performance of full scale reverse osmosis desalination plant. The effects of feed temperature, feed TDS, operating time, clean in place on feed pressure, differential pressure, permeate TDS were considered. The originality of this study lies on the application of ANN and TM to a full-scale RO desalination plant and the comparison of the two approaches in terms of accuracy and implementation feasibility. To the best of our knowledge, this is the first report for the analysis of long-term RO plant operation data using these two methods, which will provide insight into the efficient operation of full-scale RO processes.

2. Material and Methods

2.1. Full Scale RO Plant

The targeted full-scale RO plant in this study is comprised of intake, dissolved air flotation (DAF), gravity-type dual media filter (DMF), cartridge filter (CF) and RO processes. RO elements were configured in two pass. The RO plant was operated in a constant flow mode with a recovery of 34 to 37%. The plant was automatically operated and the operating data such as feed flow rate, feed pressure, feed TDS, permeate flow rate, permeate TDS and temperature were collected using a computer. The effects of fouling on the full-scale RO plant were analyzed in terms of the feed pressure, differential pressure, and permeate TDS.

2.2. ANN Model

In this study, ANN model with a perceptron network (MLP) that had a back-propagation training algorithm was used to analyze the performance of the RO process The Neural Network Toolbox V9.0 of MATLAB mathematical software was used to develop ANN model. ANN is computational models able to simulate the processing and learning functions of a human brain [15, 17]. In accordance with the human brain, an ANN model is composed of simple elements operating in parallel [18]. Neurons in a certain layer of the ANN are connected to those from the previous layer by a number

of weighted connections. In addition, there is an extra weight, named bias, which is summed to the rest of input weights [17].

The training of ANN model is carried out by adjusting the connection value among elements in order to minimize its performance factor defined as the mean squared error (MSE). Inputs are represented by x_1, x_2 and x_i, and the output by y_j. There may be many input signals to a node. The node manipulates the inputs to give a single output signal. The strength of each connection, referred to its connection weight, determines the intensity of the input signal as registered by the artificial neuron. Input data are presented to the network through the input layer, the values of which are denoted by x_i. Every input is multiplied by its corresponding weight and the node uses summation of these weighted inputs ($W_{ij} \times x_1$) to estimate an output signal using a transfer function. These weighted inputs are then summed and added to a threshold value (θ_j) to produce the node input (I_j) as shown in the equation below [17, 18]:

$$I_j = \sum_{i=1}^{n}(W_{ij} \times x_i) + \theta_j \qquad (1)$$

After a neuron performs its function, it passes its output to all of the neurons in the layer below it, providing a feed-forward path to the output through a transfer function [17]. The transfer function is one of fundamental elements, which correlate inputs with outputs in various ways based on different transfer functions. The sigmoid transfer function, the linear transfer function and the hard limit transfer function are used commonly as transfer function [17, 18]. Selecting the appropriate transfer function, algorithm and inputs can avoid the stuck in local minima and improve the generalization performance in a certain extent [19].

2.3. Tree Model

In this study, M5P tree model was used to analyze the performance of the RO process from the full-scale plant data. M5P tree model is machine-learning method which is a combination of data classification and regression [15]. The M5P tree model was created in Weka software developed at the university of Waikato, New Zealand. A comprehensive of data preprocessing and modeling technique are contained in Weka software [20]. The MP5 tree model is used for numeric prediction and each leaf stores a linear regression model that predicts the target results. The MP5 tree model is constructed in a top down way [21]. To build a M5 tree, dividing criteria must be defined. Dividing the branches criteria is based on standard deviation of the attribute values. The attribute that reduces expected error is chosen as the root of the tree. The formula of standard deviation reduction (SDR) is calculated as [20].

$$\text{SDR} = sd(T) - \sum \left[\frac{T}{T_i}\right] \times sd(T_i) \qquad (2)$$

$$sd(T) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(T_i - \bar{T})} \qquad (3)$$

Where,

T : set of attributes values

T_i : attribute value that taken from divided node according to selected attribute

$\bar{T}$: average value of the sets of T attribute

$sd(T)$: standard deviation of T

After building the tree, pruning the tree must be done to increase classification performance. When the building ends, pruning starts from the leafs to root. After every pruning, most successful tree is determined [22].

3. Results and Discussion

3.1. Feed Water Qualities and Plant Operating Data

All plant operation data and water quality parameters were normalized with plant design parameters in this study. Water quality parameters for the feed water to the RO full-scale plant were continuously monitored as shown in Fig. 1. The data from the RO plant were collected for 23 m, during which cleaning in place (CIP) was conducted two times. The plant design value of feed TDS and feed temperature was set to 44,000 mg/L and 15°C, respectively. As it is shown in Fig. 1(a), feed water TDS remained quite stable (variation = 7.0%) over the operating time. Otherwise, feed water temperature (Fig. 1(b)) changed drastically in the whole operating period (variation = 140%). The maximum and minimum relative temperatures were 1.0 and 2.5, respectively, and the average value was 1.92. The relative feed TDS values ranged from 0.91 to 0.98.

Fig. 2 shows the operational data of the RO full-scale plant. The changes in feed flowrate, feed pressure, differential pressure and permeate TDS were shown as a function of the operation
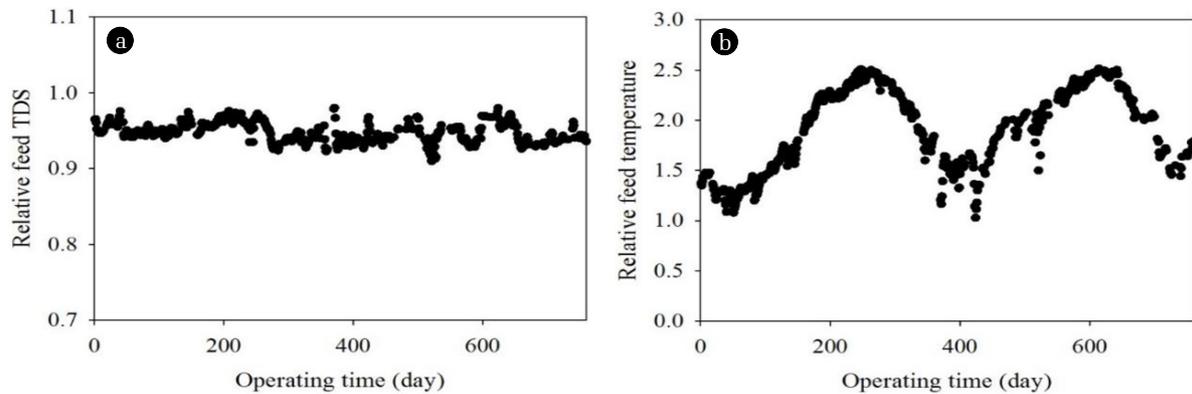

Fig. 1. Changes in feed quality parameters with operation time (a) Relative feed TDS (b) Relative feed temperature.
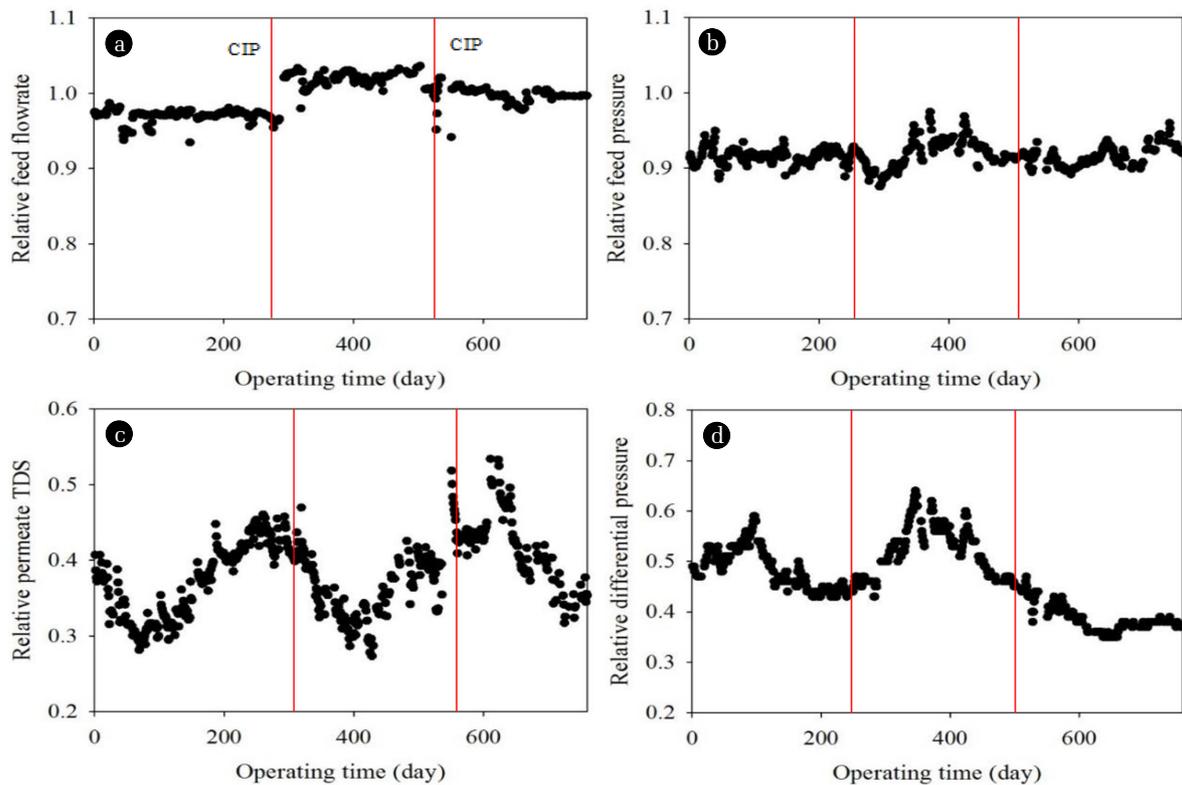

Fig. 2. Changes in operating data with operation time (a) Relative feed flow rate (b) Relative feed pressure (c) Relative permeate TDS (d) Relative differential pressure.

time. Since the plant was operated at the constant flux, the permeate flowrate remained stable with time. As shown Fig. 2(a), the feed flow also did not significantly change (variation = 11.0%). The relative feed flow rate values ranged from 0.94 to 0.104. Similar to relative feed flow, relative feed pressure did not change greatly (variation = 13.0%). The relative feed flow rate values ranged from 0.87 to 0.98. Contrarily, relative permeate TDS and relative differential pressure changed dramatically shown as Fig. 2(c), and Fig. 2(d). The relative permeate TDS significantly depended on the feed temperature. The high temperature increased the salt passage of RO membrane and thus increased the permeate TDS under constant flux and recovery operating conditions. The relative permeate TDS values ranged from 0.27 to 0.53. The relative differential pressure also depended on the feed temperature due to viscosity. The low temperature increased the viscosity of feed water and thus increased pressure drop in the pressure vessel. The relative differential pressure values ranged from 0.35 to 0.64.

The main parameters that indicate a need to CIP were increase in differential pressure between feed and concentrate and operating time in this plant. Although relative differential pressure changed dramatically, it did not exceed the CIP criteria. Therefore, the CIP was conducted once every 8 m regularly.

3.2. ANN model Set-up, Calibration and Validation

The ANN model generated to modeling the performance of the RO plant in terms of feed pressure, differential pressure and per-meates TDS. The input parameters of ANN model were carefully selected to include physically meaningful and easy-to-measure membrane operations data. The ANN model used the input parameters for the operating conditions (operating time, operating time after CIP and feed flow rate) and the feed water quality (feed TDS and feed temperature) to elevate the ANN model performance.

In this study, the ANN model is composed of 5 inputs, one hidden layer containing 10 neurons, respectively, and 1 output, respectively for prediction of relative feed pressure; relative differential pressure and relative permeate TDS. The Levenberg-Marquardt training function was selected. The experimental data was divided into three sets. 70% of experimental data are applied for training the ANN model, 15% of experimental data are applied for validation to provide an unbiased evaluation of a model fit on the training dataset while tuning model parameters, and to halt training when generalization stops improving. And the last 15% are applied to provide an unbiased test of a final model fit on the training dataset.

Fig. 3 shows the results of ANN application to the relative feed pressure, relative differential pressure and relative permeate TDS data in the full-scale RO plant. The results from the ANN model matched the experimental values very well. This suggests that the ANN model having the five inputs can successfully fit the operation data from the full-scale RO plant. No additional inputs were needed in this case.

Fig. 4. shows the *MSE* and the number of iterations. A sharp drop in the *MSE* in the first a few iterations is shown. The training
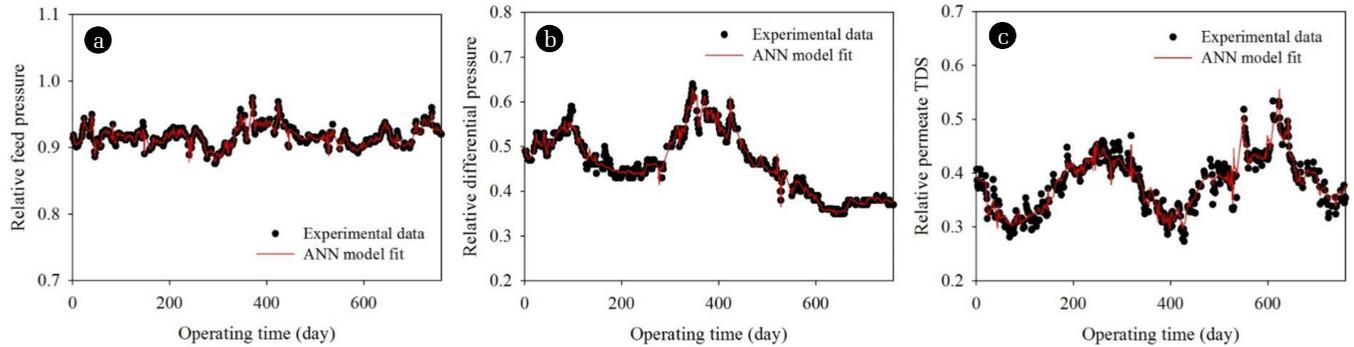

Fig. 3. Comparison of experimental data of the full-scale RO plant with the ANN model results (a) Relative feed pressure (b) Relative differential pressure (c) Relative permeate TDS.
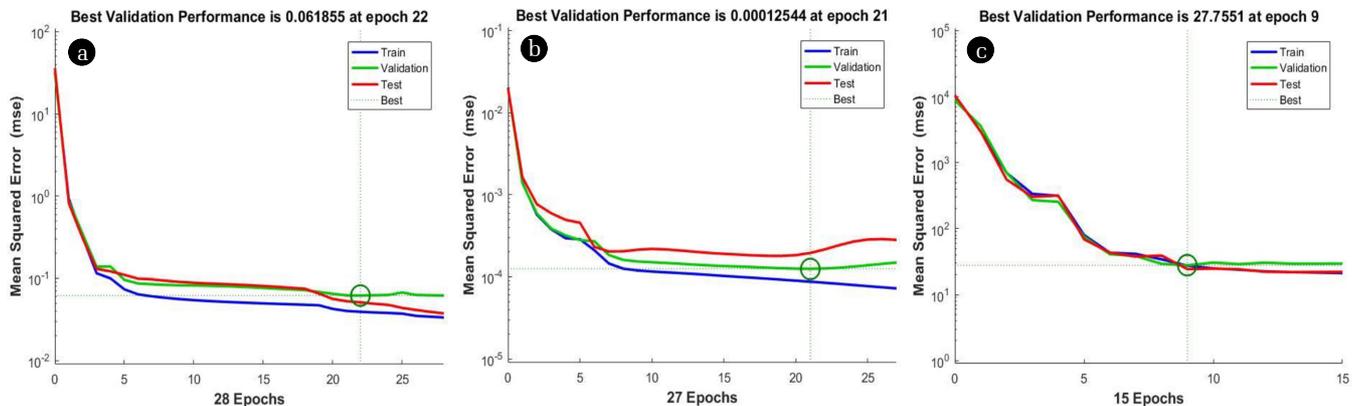

Fig. 4. MSE as a function of the number of iterations (a) Relative feed pressure (b) Relative differential pressure (c) Relative permeate TDS.

cycles stopped after 28, 27 and 15 iterations to preventing over-fitting, with a smallest validation *MSE* value of 0.00618, 0.00012 and 27.7551 at 22, 21 and 9 iterations, respectively for prediction of relative feed pressure, relative differential pressure, and relative permeate TDS.

3.3. TM Model Set-up, Calibration and Validation

The M5P tree model generated using Weka with 15% split validation for independent test of tree model. The M5P tree model, unlike many commonly statistical models such as ANN model, can explain the results and make hidden patterns known from the experimental data [23]. Same as ANN model, operating time (OT), operating time after CIP, relative feed TDS, relative feed temperature and relative feed flow rate were used as input to induce the tree model. The tree model was composed of 35, 48 and 22 multivariable linear model equations (LM) to predict relative feed pressure, relative differential pressure and relative permeate TDS, respectively. Fig. 5 shows the structure of tree model to predict the relative feed pressure. The input parameter at the top of the tree structure was the most particular [24]. In this study, temper-
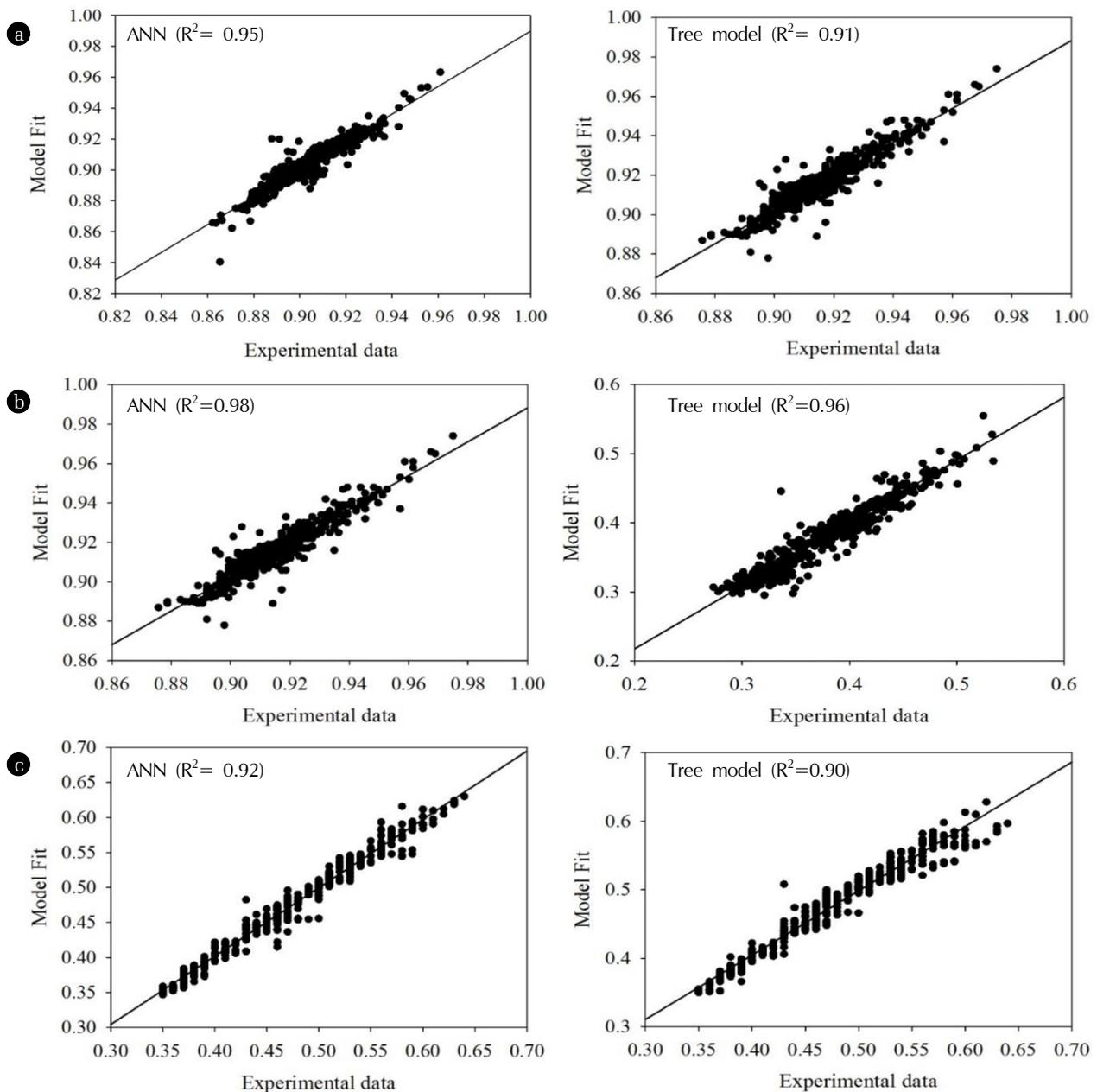

Fig. 7. Comparison of model predictions with full-scale RO plant data (a) Relative feed pressure (b) Relative differential pressure (c) Relative permeate TDS.

Table 1. Summary of the Statistical Parameters Obtained from the Model Fits

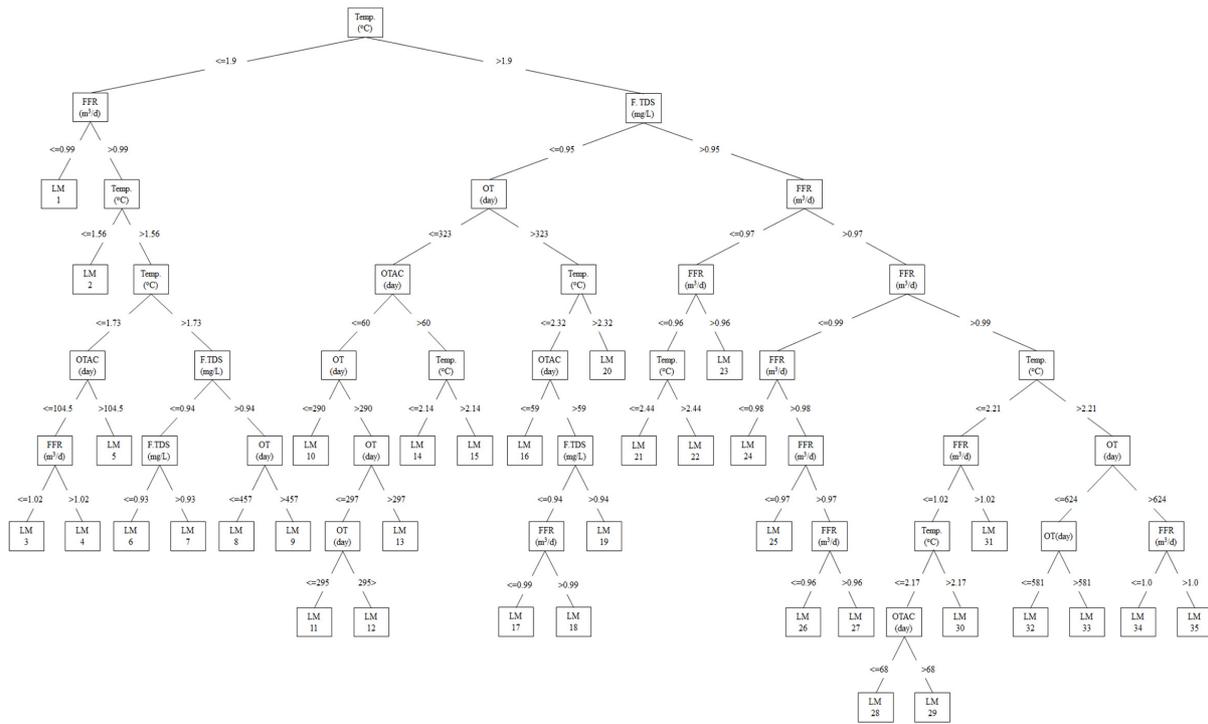| Model	Items		Relative feed pressure	Relative differential pressure	Relative permeate TDS
ANN	Standard deviation (SD)	Experimental data	0.014	0.067	0.053
		Model Fit	0.014	0.065	0.052
	Standard deviation ratio (SDR)	Experimental data	1.589	14.189	13.810
		Model Fit	1.507	13.728	13.569
	Coefficient of determination (R^2)		0.95	0.98	0.92
Model Tree	Standard deviation (SD)	Experimental data	0.014	0.067	0.053
		Model Fit	0.015	0.067	0.054
	Standard deviation ratio (SDR)	Experimental data	1.589	14.189	13.810
		Model Fit	1.552	13.988	13.783
	Coefficient of determination (R2)		0.91	0.95	0.9

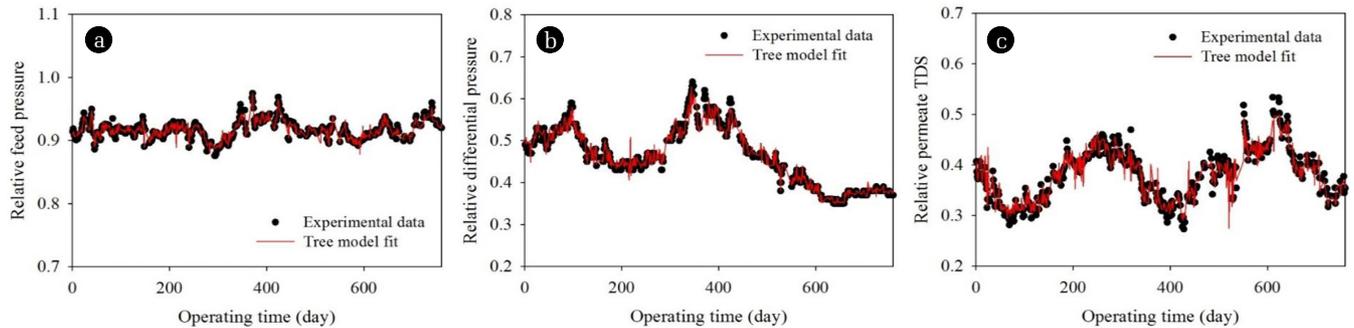**Fig. 5.** M5P tree for relative feed pressure of full-scale RO plant.

Fig. 6. Comparison of experimental data of the full-scale RO plant operation with the tree model predictions (a) Relative feed pressure (b) Relative differential pressure (c) Relative permeate TDS.

ature was the most influential parameter affecting the output in the M5P tree mode. Each LM enables the estimation of the relative feed pressure; relative differential pressure and relative permeate TDS as a linear regression of multiple operating parameters. For example, LM 1 for prediction of relative feed pressure is valid when the relative temperature is below 1.9 and the relative feed flowrate is under 0.98. The LM equations of tree model can notify the most influential parameters because each parameter is multiplied by a weighting factor [24].

The comparison between measured and modeled relative feed pressure, relative differential pressure, and relative permeate TDS values are shown in Fig. 6. It is observed that the model results track the observed data very well. The tree model showed high strength and a linear relationship direction between the model and experimental data.

3.3. Comparison of the ANN Model and Tree Model

The results of the ANN model and tree model were evaluated based on the comparison between the model output and the experimental data using coefficient of determination, R^2 as shown in Fig. 7. The statistical parameters obtained from the model fits are also summarized in Table 1. The ANN model and tree model resulted in good agreements with the plant data. It is observed that the output tracks the targets very well for relative feed pressure [R^2 value = 0.95 (ANN), 0.91 (TM)], validation relative differential pressure [R^2 value = 0.98 (ANN), 0.96 (TM)], and relative permeate TDS [(R^2 value = 0.92 (ANN), 0.90 (TM)] as shown in Fig. 7. This indicates that both ANN and TM have the potential for long-term prediction of the RO performance in full-scale desalination plant.

In fact, ANN showed a slightly higher R^2 value than TM. Considering the complexity of ANN for the implementation in full-scale plants, however, it appears that TM is more appropriate than ANN. As shown in Fig. 5, the result of TM is a logic that can be relatively easily reflected on the operating system for RO process. On the other hand, the result of ANN cannot be directly used by the RO operating systems.

4. Conclusions

In this work, the performance of both ANN model and tree model (TM) was evaluated to analyze the effects of operating conditions and the feed water quality on the feed pressure, differential pressure, and permeate TDS of a full scale RO plant. The following conclusions were withdrawn:

Both the ANN and tree model demonstrated a high strength and direction with respect to the linear relationship between the simulated and observed data. This implies that the two models can be useful for analyzing the performance of full-scale RO plants.

In terms of the model fit accuracy, the ANN model was slightly better than the tree model. However, the difference in R^2 values between the ANN and the tree model was not significant, suggesting that the both models are applicable to fit the plant data.

In terms of the practical implementation of these models in a full scale RO plant, it appears that the tree model has advantages over the ANN model. This is because the ANN model is a "black-box"

approach and thus does not intuitively show how it can predict the data. On the other hand, the tree model is expressed as a set of simple functions. Therefore, tree model seems to be more adequate than ANN for it application to a full-scale RO plant.

Acknowledgment

Author Contributions

Y. L. (Ph.D.) and K. S. (Ph.D.) conducted all the experiments. Y.P. (Ph.D. student) conducted the ANN modeling. Y.C. (Ph.D.) and S.L. (Professor) wrote the manuscript and conducted all modeling.

References

1. Water crises are a top global risk. World Economic Forum. 16 January 2015. Retrieved 30 December 2017.
2. NATURE-BASED SOLUTIONS FOR WATE. The United Nations World Water Development Report 2018.
3. Sadek A. Water desalination: An imperative measure for water security in Egypt. *Desalination* 2010;250:876-884.
4. Sukitpaneenit P, Chung TS. High performance thin-film composite forward osmosis hollow fiber membranes with macro-void-free and highly porous structure for sustainable water production. *Environ. Sci. Technol.* 2012;46:7358-7365.
5. Kang G, Cao Y. Development of antifouling reverse osmosis membranes for water treatment: A review. *Water Res.* 2012;46:584-600.
6. Goosen MFA, Sablani SS, Al-Hinai H, Al-Obeidani S, Al-Belushi R, Jackson D. Fouling of reverse osmosis and ultrafiltration membranes: A critical review. *Separation Sci. Technol.* 2011;141:269-289.
7. Koltuniewicz A., Noworyta A. Dynamic properties of ultra-filtration systems in light of the surface renewal theory. *Ind. Eng. Chem. Res.* 1994;33:1771-1779.
8. Guo W, Ngo HH, Li J. A mini-review on membrane fouling. *Bioresour. Technol.* 2012;122:27-34.
9. Roehl EA, Ladner DA, Daamen RC, et al. Modeling fouling in a large RO system with artificial neural networks. *J. Membr. Sci.* 2018;552:95-106.
10. Tang CY, Chong TH, Fane AG. Colloidal interactions and fouling of NF and RO membranes: A review. *Adv.Colloid. Interf. Sci.* 2011;164:126-143.
11. Malaeb L, Ayoub GM. Reverse osmosis technology for water treatment: State of the art review. *Desalination* 2011;267:1-8.
12. Choi JS, Kim JT. Modeling of full-scale reverse osmosis desalination system: Influence of operational parameters. *J. Ind. Eng. Chem.* 2014:21:261-268

13. Ahmed AA, Robert WL. Fouling strategies and the cleaning system of NF membranes and factors affecting cleaning efficiency. *J. Membr. Sci.* 2007:303:4–28

14. Soleimani R, Shoushtari NA, Mirza B, Salahi A. Experimental nvestigation, modeling and optimization of membrane separation using artificial neural network and multi-objective optimization using genetic algorithm. *Chem. Eng. Res. Des.* 2013;91:883-903.

15. Dimitri PS, Khada ND. Model trees as an alternative to neural networks in rainfall-runoff modelling 2003:48:399-411.

16. Quinlan JR. Learning with continuous classes. Proceedings 5th Australian Joint Conference on Artificial Intelligence, World Scientific Press, Singapore 1992:343-348.

17. Delgrange VN, Cabassud N, Cabassud M, Durand-Bourlier L, Laine JM. Neural networks for prediction of ultrafiltration trans-membrane pressure: Application to drinking water production. *J. Membr. Sci. 1*998;150:111-123.

18. Schmitt F, Banu R, Yeom IT, Do KU. Development of artificial neural networks to predict membrane fouling in an anoxic-aero-bic membrane bioreactor treating domestic wastewater. *Biochem. Eng. J.* 2018;133:47-58.

19. Liu QF, Kim SH, Lee S. Prediction of microfiltration membrane fouling using artificial neural network models. *Sep. Purif. Technol.* 2019;70:96-102.

20. Witten HI, Frank E, Hall MA. Data mining, principle machine learning tools and techniques. Burlingtan: Elsevier; 2011.

21. Wang Y, Witten HI. Inducing model trees for continuous classes. In: Proceedings of the 9th European Conference on Machine Learning; 1997; Prague.

22. Wu X, Kumar V. CART. Classification and regression trees, top ten algorithms in data mining. Boca Raton: CRC Press; 2009.

23. Onyari EK, Ilunga FM. Application of MLP neural network and M5P model tree in predicting streamflow: A case study of Luvuvhu catchment, South Africa. *Int. J. Innov. Manage. Technol*. 2013;4:1-15.

24. Dalmau M, Atanasova N, Gabarrón S, Roda IR, Comas J. Comparison of a deterministic and a data driven model to describe MBR fouling. *Chem. Eng. J.* 2015;260:300-308.