Time series modeling of biochemical oxygen demand at the upstream catchment of Feitsui Reservoir, Taiwan

Yung-Chia Chiu, Chih-Wei Chiang and Tsung-Yu Lee

ABSTRACT

The adaptive neuro fuzzy inference system (ANFIS) has been proposed to model the time series of water quality data in this study. The biochemical oxygen demand data collected at the upstream catchment of Feitsui Reservoir in Taiwan for more than 20 years are selected as the target water quality variable. The classical statistical technique of the Box-Jenkins method is applied for the selection of appropriate input variables and data pre-processing of using differencing is implemented during the model development. The time series data obtained by ANFIS models are compared to those obtained by autoregressive integrated moving average (ARIMA) and artificial neural networks (ANNs). The results show that the ANFIS model identified at each sampling station is superior to the respective ARIMA and ANN models. The R values at all sampling stations of the training and testing datasets are 0.83–0.98 and 0.81–0.89, respectively, except at Huang-ju-pi-liao station. ANFIS models can provide accurate predictions for complex hydrological processes, and can be extended to other areas to improve the understanding of river pollution trends. The procedure of input selection and the pre-processing of input data proposed in this study can stimulate the usage of ANFIS in other related studies.

Key words | adaptive neuro fuzzy inference system, artificial neural networks, autoregressive integrated moving average, biochemical oxygen demand, water quality

INTRODUCTION

It is important to understand surface water quality due to its effect on aquatic ecosystems and public health. The common metric for assessing an aquatic ecosystem’s organic pollution level is the biochemical oxygen demand (BOD). Measuring the BOD serves as an indicator of the health of the aquatic system. The difficulty with measuring BOD is that after the initial sample is obtained, it requires 5 days to complete. The final analysis of the results is done on the fifth day and the final result is referred to as BOD5 (Delzer & McKenzie 1999). Due to this time lag and high financial expense of generating field BOD5 measurements, the estimation of BOD5 has led to active research in mathematical models (Roider & Adrian 2007). Mathematical models are based on the underlying physics of the environmental system or approximations of it. One major limitation is that the models can be sensitive to parameter changes that lead to different predictions (Chau 2006).

An alternative to mathematical modeling is time series modeling. This approach has become an area of active research due to its applicability in a variety of fields such as dynamic systems, nonlinear signal processing, pattern recognition, and hydrology. To construct a time series model, a set of historical observations are combined to extrapolate a forecast of future measurements. The advantage of this modeling approach is that it does not rely on the underlying physics of an environmental system, but solely uses the historical data. This approach is also useful when there is a difficulty in collecting all explanatory variables continuously. For modeling water quality, the time series methods of autoregression (AR) and autoregressive moving average (ARMA) have

Yung-Chia Chiu (corresponding author)
Chih-Wei Chiang
Institute of Applied Geosciences,
National Taiwan Ocean University,
Keelung,
Taiwan
E-mail: ycchiu@mail.ntou.edu.tw
Tsung-Yu Lee
Department of Geography,
National Taiwan Normal University,
Taipei,
Taiwan
been successfully applied (Tiwari et al. 1996; Cai & Tiwari 2000; Schoch et al. 2009; Abaurrea et al. 2011). More recently a more generalized version of ARMA called autoregressive integrated moving average (ARIMA) has been applied to water quality models (Ahmad et al. 2001; Kurunc et al. 2005; Zou et al. 2009; Abdu et al. 2012). An ARIMA model is particularly applied in some cases where the time series data show evidence of non-stationarity features. Although ARIMA models are quite flexible in that they can represent several different types of time series, their major limitation is the pre-assumed linear form of the model and may not be adequate for complex nonlinear real world problems (Oliveira-Esquerre et al. 2004).

In order to overcome the assumption of linearity, time series have been incorporating various artificial intelligence (AI) techniques (Zhang et al. 1998). Among various AIs, artificial neural networks (ANNs) have been found to be well suited to modeling a system for which an explicit understanding of the complex input and output relationship is unknown. ANNs have been widely applied in engineering fields, and successfully used in hydrological processes, water resources, evaporation, flood forecasting, reservoir operations, and recorded weather data (Hsu et al. 1995; Wen & Lee 1998; Aguilera et al. 2001; Oliveira-Esquerre et al. 2004; Kisi 2008; Gao et al. 2010; Chen et al. 2014; Shiri et al. 2014). ANNs have been also extended to forecast water quality parameters and estimate nutrient concentration from pollution sources of watersheds (Maier & Dandy 1996; Suen & Eheart 2003; Sengorur et al. 2006; Singh et al. 2009; Basant et al. 2010; Ay & Kisi 2012; Kisi et al. 2013; Marti et al. 2013; Verma & Singh 2013; Wen et al. 2013). Maier et al. (2010) reviewed 210 papers on the prediction of water quantity and quality variables in rivers and provided a taxonomy of methods used for the development of ANN models for river systems.

Fuzzy logic theory based on the linguistic uncertainty expression rather than numerical uncertainty was first developed by Zadeh (1965) and has become popular to various engineering problems and hydrological modeling since the late 1990s (Jang et al. 1997). Recently, adaptive neuro fuzzy inference system (ANFIS), which consists of the ANN and fuzzy logic methods, has received attention in database management, system design, and water resources planning and forecasting. Nayak et al. (2005) stated that while ANFIS models do not provide information on the hydrologic process physics, they are still useful for time series forecasting where the main concern is accurate predictions of a variable at specific watershed locations. ANFIS has been successfully used for prediction stream flow (Chang & Chen 2001; Nayak et al. 2004; Firat & Gungor 2007; Firat & Gungor 2008), reservoir inflow or water level (Chang & Chang 2006; Lohani et al. 2012), flood forecast (Nayak et al. 2005; Chen et al. 2006), sediment yield (Firat & Gungor 2010), and evaporation/transpiration (Shiri et al. 2011, 2015a, 2015b, 2015; Kisi et al. 2012; Pour Ali Baba et al. 2015) for many years.

Although the ANFIS method has been certified as a powerful tool for prediction and plenty of applications have been published (primarily in water quantity), there is little discussion in the literature of using this hybrid computing system to model the time series of water quality, especially BOD₅ (Maier et al. 2010). Besides, there is no systematic procedure for the development of ANFIS (or ANN) models, nor a comprehensive study to compare different modeling techniques (Wu et al. 2014). The main purpose of this study is to present ANFIS modeling of BOD₅ as time series data. This study further provides a systematic procedure of input selection and the pre-processing of the input data, both of which are embedded in the model development. To verify the application of this approach, the upstream catchment of Feitsui Reservoir located at the northeast part of Taiwan is chosen as the case study area. A large amount of BOD₅ data in terms of space and time is collected in this area and used to develop the models. The capabilities of proposed ANFIS models are compared to ARIMA and ANN models.

**METHODOLOGY**

In this section, the basic concepts and modeling process of proposed ARIMA, ANN, and ANFIS models are briefly reviewed.

**Box-Jenkins autoregressive integrated moving average (ARIMA)**

In an ARIMA \((p,d,q)\) model, the predictions of the future values are constrained to be linear functions of past
observations and random errors. The underlying process that generates the time series with the mean $\mu$ has the form:

$$\phi(B)\nabla^d(y_t - \mu) = \theta(B)a_t$$

(1)

where $t$ is the time index, and $y_t$ is the actual value, and $a_t$ is a random error variable. $\phi(B)$ and $\theta(B)$ are the autoregressive and moving average operators, respectively, and can be written as:

$$\phi(B) = 1 - \sum_{i=1}^{p} \phi_i B^i$$

(2)

$$\theta(B) = 1 - \sum_{j=1}^{q} \theta_j B^j$$

(3)

where $\phi_j (i = 1, 2, \ldots, p)$ and $\theta_j (j = 1, 2, \ldots, q)$ are model parameters. $\nabla = (1 - B)$ and $B$ is the backward shift operator. $p$ and $q$ are integers greater than or equal to zero and referred to as the orders of the autoregressive and moving average parts of the model, respectively. $d$ is also an integer and referred to as the order of differencing. Random errors, $a_t$, are assumed to be independently and identically distributed (i.i.d.) with a mean of zero and a constant variance of $\sigma^2$.

ARIMA models form an important part of the Box-Jenkins method (Box & Jenkins 1976) to time series modeling. By matching the empirical autocorrelation patterns with the theoretical ones, Box & Jenkins (1976) proposed to compute and examine the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the time series data to identify the orders of $p$ and $q$ in ARIMA. The detailed mathematical background and the steps of model development can be found in Box & Jenkins (1976). In this study, the ARIMA model is created for each sampling station by the Box-Jenkins method and its basic structure and parameters are identified.

Artificial neural networks (ANNs)

In this study, a multilayer perceptron neural network was chosen to model time series data of BOD$_5$. The selected network has three layers, i.e., one input layer with varied input variables, one hidden layer with varied nodes, and one output layer with one output variable. The output of a neuron can be expressed as:

$$y_i = f_2\left( \sum_{j=1}^{I} w_{ij}^{(2)} x_j + b_i \right)$$

(4)

where $x_i$ is the input variable, $w_{ij}^{(2)}$ and $w_{jk}^{(2)}$ are the interconnection weights between input and hidden layers, and hidden and output layers, respectively. $b_i$ and $b_k$ are bias values for hidden and output layers, respectively. $f_1(\cdot)$ and $f_2(\cdot)$ are the activation functions for hidden and output layers, respectively.

The back-propagation algorithm introduced by Rumelhart et al. (1986) is perhaps the most popular algorithm for training ANNs. With the back propagation algorithm, the linear (Equation (5)), logistic (Equation (6)), and tangent (Equation (7)) functions are the most commonly used activation functions in the construction of ANNs,

purelin: $f(x) = x$  

(5)

logsig: $f(x) = \frac{1}{1 + e^{-x}}$  

(6)

tansig: $f(x) = \frac{2}{1 + e^{-2x}} - 1$  

(7)

The training of the neural network embraces the choices of architecture, activation functions, training algorithms, and network parameters, the testing of the trained network, and the usage of trained neural network for simulation and prediction. All these steps are adopted to develop the ANN models in this study. The three-layer feed-forward ANN model trained by the back-propagation algorithm is developed for each station and the determination of input variables is based on the analysis obtained from Box-Jenkins ARIMA.

Adaptive neuro fuzzy inference system (ANFIS)

Figure 1 presents the five-layer structure of ANFIS with two inputs variables. In this study, Takagi-Sugeno fuzzy inference system (FIS) is used for the time series modeling of
BOD$_5$ (Takagi & Sugeno 1985). For the first-order Takagi-Sugeno FIS, two typical rules can be expressed as:

Rule 1: If $x_1$ is $A_1$ and $x_2$ is $B_1$, then $f_1 = p_1 x_1 + q_1 x_2 + r_1$

Rule 2: If $x_1$ is $A_2$ and $x_2$ is $B_2$, then $f_2 = p_2 x_1 + q_2 x_2 + r_2$

where $x_1$ and $x_2$ are the inputs, $A_i$ and $B_i$, which are the linguistic labels, $f_i$ is the output function, $p_i$, $q_i$, and $r_i$ are the consequent parameters (output function parameters).

Layer 1: each node in this layer generates membership grades of the crisp inputs and each node’s output $O_i^1$ is calculated by:

$$Q_i^1 = \mu_{A_i}(x_i) \quad \text{for} \quad i = 1, 2$$

$$Q_i^1 = \mu_{B_i}(x_i) \quad \text{for} \quad i = 3, 4$$

where $O_i^1$ denotes the output of the $i$th node in the first layer, and $\mu_{A_i}$ and $\mu_{B_i}$ are the membership functions (MF) for $A_i$ and $B_i$ linguistic labels, respectively. The MF can be any appropriate functions that are continuous and piecewise differentiable such as Gaussian, bell-shaped, trapezoidal-shaped, and triangular-shaped functions. For example, the bell-shaped function can be expressed as follows:

$$\mu_{A_i}(x_i) = \frac{1}{1 + \left(\frac{x_i - c_i}{a_i}\right)^2}$$

where $a_i$, $b_i$, and $c_i$ are the premise parameter set and with the change of these parameters the bell-shaped function varies accordingly.

Layer 2: every node in this layer multiplies the incoming signals, and the output represents the firing strengths of a rule, $O_i^2$ or $w_i$, is computed as:

$$Q_i^2 = w_i = \mu_{A_i}(x_1)\mu_{B_i}(x_2), \quad i = 1, 2$$

Layer 3: this layer calculates the ratio of the $i$th rule’s firing strength to the sum of all rule’s firing strengths. It normalizes the firing strength and is calculated as:

$$Q_i^3 = \bar{w}_i = \frac{w_i}{\sum_i w_i}, \quad i = 1, 2$$

Layer 4: this layer calculates only the sum of the signals of the third and second layers of the network,

$$Q_i^4 = \bar{f}_i = \bar{w}_i(p_1 x_1 + q_1 x_2 + r_1), \quad i = 1, 2$$

Layer 5: this layer is called the output nodes in which the single node computes the overall output by summing all incoming signals:

$$Q_i^5 = \sum_i \bar{w}_i \bar{f}_i = \sum_i \frac{w_i p_i x_1 + q_i x_2 + r_i}{\sum_i w_i}$$

A hybrid-learning algorithm which combines the gradient descent and least-squares methods to identify parameters defined in the adaptive networks has been proposed. The detailed mathematical background and hybrid-learning algorithm can be found in Jang (1993), Jang et al. (1997) and Nayak et al. (2004). In this study, the five-layer ANFIS model trained by the hybrid-learning algorithm is developed for each station, and the determination of input variables is also based on the analysis obtained from Box-Jenkins ARIMA.

Performance criteria

To determine the performance of each identified ARIMA, ANN, and ANFIS models, four different criteria are used: the Pearson coefficient of correlation ($R$), the root mean square error (RMSE), the mean absolute percentage error (MAPE), and the mean absolute error (MAE). Each criterion can be computed as:

$$R = \frac{\sum_{t=1}^{T_0} (\hat{y}_t - \bar{y})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^{T_0} (\hat{y}_t - \bar{y})^2} \sqrt{\sum_{t=1}^{T_0} (y_t - \bar{y})^2}}$$
\[ \text{RMSE} = \sqrt{\frac{1}{T_0} \sum_{t=1}^{T_0} (y_t - \hat{y}_t)^2} \quad (15) \]

\[ \text{MAPE} = \frac{1}{T_0} \sum_{t=1}^{T_0} \frac{|y_t - \hat{y}_t|}{y_t} \times 100 \quad (16) \]

\[ \text{MAE} = \frac{1}{T_0} \sum_{t=1}^{T_0} |y_t - \hat{y}_t| \quad (17) \]

where \( y_t \) and \( \hat{y}_t \) denote the target (measured) and network output (simulated) values at time period \( t \), respectively. \( \bar{y} \) and \( \bar{\hat{y}} \) represent the averages of \( y_t \) and \( \hat{y}_t \), respectively. \( T_0 \) denotes the number of measurements. Higher value of \( R \) and smaller values of RMSE, MAPE, and MAE ensure better performance of the model.

**STUDY AREA AND BIOCHEMICAL OXYGEN DEMAND DATA**

In the case study considered, ARIMA, ANNs, and ANFIS were used to model time series data of \( \text{BOD}_5 \) at the upstream catchment of Feitsui Reservoir in Taiwan. Feitsui Reservoir, located in the northern part of Taiwan (Figure 2), is a water-supply reservoir for the Taipei metropolitan area with a population of over five million. The...
main dam was built in 1987 and located downstream of Peishih Creek. The reservoir is mainly recharged by the Peishih Creek, which has two main tributary creeks, Jingua-liau and Dai-yu-ku Creeks (Figure 2). Due to the excessive pollutant loads from municipal, industrial wastewater, and agricultural run-off at the upstreams, there has been a trend of water quality decline in the reservoir (Taipei Water Management Committee, 2003). According to the Carlson trophic state index for water quality of the reservoir, the water quality condition in the Feitsui Reservoir was in the range of mesotrophic and eutrophic (Kuo et al., 2003).

The time series data of BOD₅ used in this study were generated through continuous monitoring of water quality at the upstream catchment of Feitsui Reservoir from 1986 to 2012. The data were monitored monthly at eight different sampling stations that cover most of the upstream catchment. Three stations, Ping-lin, Dai-yu-ku Creek, and Jingua-liao Creek (Figure 2), are managed by Taipei Feitusi Reservoir Administration, and the total number of measurements is 972 from 1/1986 to 12/2012. Five stations, Kuo-lai, Bi-hu, Shui-yuan Bridge, Da-lin Bridge, and Huang-ju-pi-liao (Figure 2), are managed by Taipei Water Management Committee and the total number of measurements is 1345 from 1/1987 to 5/2009.

Table 1 presents the basic statistics of the measured BOD₅ concentrations for each of the sampling stations. The variation in BOD₅ concentration corresponds to the nature and types (point and non-point) of sources distributed in the large geographical area of watershed. The larger standard deviation (SD) or coefficient of variation (CV) of BOD₅ may be directly related to more industrial, agricultural, and anthropogenic activities at the downstream catchment. In the upper-course (Kuo-lai and Bi-hu stations), the river water quality is mainly dominated by the variables of geogenic origin, and there are no major pollution sources in the region. In the mid- or lower-course (Jingua-liao Creek, Shui-yuan Bridge, and Huang-ju-pi-liao stations), the river receives a heavy load of untreated wastewater from nearby townships and is dominated by the variables of anthropogenic origin.

When developing ARIMA, ANN, and ANFIS models, only the time series data of BOD₅ were used to develop the models and no other water quality data were considered, such as pH, temperature, turbidity, electrical conductivity, or suspended solids. The entire dataset for each station is divided into two sets, the training and testing sets. Seventy percent of the data is used for training, and 30% is used for testing. For the Ping-lin, Dai-yu-ku Creek, and Jingua-liao Creek stations, the data from 1/1986 to 11/2004 (227 measurements for each station) are used for training, and the data from 12/2004 to 12/2012 (97 measurements for each station) are used for testing. For the Kuo-lai, Bi-hu, Shui-yuan Bridge, Da-lin Bridge, and Huang-ju-pi-liao stations, the data from 1/1987 to 8/2002 (188 measurements for each station) are used for training, and the data from 9/2002 to 5/2009 (81 measurements for each station) are used for testing. The performances of all identified models are evaluated by R, RMSE, MAPE, and MAE.

<table>
<thead>
<tr>
<th>Station</th>
<th>Observation period</th>
<th>Max (mg/L)</th>
<th>Min (mg/L)</th>
<th>Mean (mg/L)</th>
<th>Median (mg/L)</th>
<th>SD (mg/L)</th>
<th>CV (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ping-lin</td>
<td>1987–2012</td>
<td>4.30</td>
<td>0.00</td>
<td>0.69</td>
<td>0.60</td>
<td>0.51</td>
<td>75</td>
</tr>
<tr>
<td>Dai-yu-ku Creek</td>
<td>1987–2012</td>
<td>3.70</td>
<td>0.00</td>
<td>0.63</td>
<td>0.50</td>
<td>0.49</td>
<td>77</td>
</tr>
<tr>
<td>Jingua-liao Creek</td>
<td>1987–2012</td>
<td>5.90</td>
<td>0.10</td>
<td>0.72</td>
<td>0.50</td>
<td>0.63</td>
<td>88</td>
</tr>
<tr>
<td>Kuo-lai</td>
<td>1987–2009</td>
<td>2.70</td>
<td>0.10</td>
<td>0.55</td>
<td>0.50</td>
<td>0.35</td>
<td>65</td>
</tr>
<tr>
<td>Bi-hu</td>
<td>1987–2009</td>
<td>2.80</td>
<td>0.10</td>
<td>0.59</td>
<td>0.50</td>
<td>0.37</td>
<td>63</td>
</tr>
<tr>
<td>Da-lin Bridge</td>
<td>1987–2009</td>
<td>2.20</td>
<td>0.10</td>
<td>0.64</td>
<td>0.50</td>
<td>0.34</td>
<td>53</td>
</tr>
<tr>
<td>Shui-yuan Bridge</td>
<td>1987–2009</td>
<td>3.40</td>
<td>0.10</td>
<td>0.86</td>
<td>0.90</td>
<td>0.55</td>
<td>63</td>
</tr>
<tr>
<td>Huang-ju-pi-liao</td>
<td>1987–2009</td>
<td>4.30</td>
<td>0.10</td>
<td>0.92</td>
<td>1.10</td>
<td>0.64</td>
<td>70</td>
</tr>
</tbody>
</table>

Max, maximum; Min, minimum; SD, standard deviation; CV, coefficient of variation.
RESULTS AND DISCUSSION

In this section, the time series data of BOD$_5$ at the upstream catchment of reservoir are used to demonstrate the developments of ARIMA, ANN, and ANFIS models.

Box-Jenkins autoregressive integrated moving average

Before creating the ARIMA model for each sampling station, the seasonality and stationarity of the time series data at each sampling station are examined. The temporal variations of measured BOD$_5$ are plotted (Figure 4) and no seasonality is observed in the measurements for all the stations. The ACF chart at each station is also examined and the correlations show that no seasonality is detected. To test if the data are stationary, the unit root test of augmented Dickey-Fuller test (ADF) (Dickey & Fuller 1979) is employed. The results of this test show that the time series data at all sampling stations are non-stationary; therefore, the first-order difference, i.e., $d = 1$, is applied to each dataset to achieve stationarity.

After determining the order of differencing, the training dataset is used to identify appropriate orders of $p$ and $q$ by examining the PACF and ACF charts, respectively. Figure 3(a) shows the ACF and PACF charts at Ping-lin station and the correlations are significant only at time lag 1 of ACF and time lags 1–7 of PACF. Consequently, seven ARIMA $(p, d, q)$ models are considered as the potential candidates to be identified and their associated $R$, RMSE, MAPE, and MAE for the training and testing datasets are calculated to determine the optimal model. The identified optimal ARIMA $(p, d, q)$ model selected from the potential candidates is ARIMA $(7, 1, 1)$ and given in the notation as follows:

Equation (18) could be rewritten in the following form and means that $y_t$ is the function of $y_{t-1}$, $y_{t-2}$, $y_{t-3}$, $y_{t-4}$, $y_{t-5}$, $y_{t-6}$, $y_{t-7}$, $y_{t-8}$, and $a_{t-1}$,

$$y_t = f(y_{t-1}, y_{t-2}, \ldots, y_{t-8}, a_{t-1})$$  \hspace{1cm} (20)

After identifying the optimal ARIMA model, it is validated by examining the autocorrelation of residuals. The residual ACF and PACF charts of BOD$_5$ at Ping-lin station are presented in Figure 3(b), and the series are all within the 95% confidence intervals. This then indicates that the residuals are an approximate white noise time series, not significant statistically, and the model is properly identified.

The identifications of ARIMA models at other sampling stations followed the same procedures that were described previously. All of the identified ARIMA models at all sampling stations are effective significantly in statistics, and their basic structures and performances are presented in Table 2. It can be seen from the table that all sampling stations do not perform well, i.e., the $R$ value is low, and the RMSE/MAPE/MAE values are high. Figure 4 displays the entire time series (including training and testing datasets) for the simulated and measured BOD$_5$ at all sampling stations. According to the results of simulation, the ARIMA models can only depict the trend of time series data of BOD$_5$, but the extreme concentrations, especially high values, are not well described. This unsatisfactory result due to the complexity of aquatic systems and abundant anthropogenic activities is expected. Although all of the information contained in the data has been extracted by the linear ARIMA model, the model has reached its capability of capturing linear relationship between inputs and output. The process variables may have some degrees of nonlinear relationship that results in the lower accuracy when using linear ARIMA model to model time series data of BOD$_5$.

Artificial neural networks

For the identification of ANNs, the original time series dataset of BOD$_5$ measured at each station is first differenced and then normalized to an interval between $–1$ and $+1$. This particular form of data pre-processing of using differencing has never been observed to our knowledge in the related
The selection of input variables (nodes), a very important aspect for the ANN modeling, is traditionally based on either an ad hoc approach or empirical methods. However, in this study, a new approach is introduced and implemented for the first time, which involves a more systematic and data-driven selection process.

Figure 3 | (a) ACF and PACF charts, (b) residual ACF and residual PACF charts of BOD5 at Ping-lin station after fitting the ARIMA (7,1,1) model.
Figure 4 | Simulated and measured BOD₅ of ARIMA models at: (a) Ping-lin; (b) Dai-yu-ku Creek; (c) Jin-gua-liau Creek; (d) Kuo-lai; (e) Bi-hu; (f) Da-lin Bridge; (g) Shui-yuan Bridge; and (h) Huang-ju-pi-liao.

Table 2 | The basic structure and performance of optimized ARIMA model for each sampling station

<table>
<thead>
<tr>
<th>Station</th>
<th>Structure</th>
<th>Dataset</th>
<th>R</th>
<th>RMSE</th>
<th>MAPE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ping-lin</td>
<td>ARIMA (7,1,1)</td>
<td>Training</td>
<td>0.3803</td>
<td>0.2720</td>
<td>84.6107</td>
<td>0.2220</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Testing</td>
<td>0.2125</td>
<td>0.5186</td>
<td>95.8707</td>
<td>0.3778</td>
</tr>
<tr>
<td>Dai-yu-ku Creek</td>
<td>ARIMA (5,1,1)</td>
<td>Training</td>
<td>0.3110</td>
<td>0.3278</td>
<td>90.2363</td>
<td>0.2592</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Testing</td>
<td>0.1336</td>
<td>0.4923</td>
<td>92.1265</td>
<td>0.3511</td>
</tr>
<tr>
<td>Jin-gua-liau Creek</td>
<td>ARIMA (5,1,1)</td>
<td>Training</td>
<td>0.4594</td>
<td>0.5069</td>
<td>77.3537</td>
<td>0.3543</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Testing</td>
<td>0.1418</td>
<td>0.5547</td>
<td>113.6521</td>
<td>0.4144</td>
</tr>
<tr>
<td>Kuo-lai</td>
<td>ARIMA (6,1,1)</td>
<td>Training</td>
<td>0.3842</td>
<td>0.3193</td>
<td>42.1655</td>
<td>0.2456</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Testing</td>
<td>0.1746</td>
<td>0.3770</td>
<td>83.1859</td>
<td>0.2589</td>
</tr>
<tr>
<td>Bi-hu</td>
<td>ARIMA (6,1,1)</td>
<td>Training</td>
<td>0.3746</td>
<td>0.2441</td>
<td>44.2257</td>
<td>0.1954</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Testing</td>
<td>0.2804</td>
<td>0.3850</td>
<td>75.6060</td>
<td>0.2675</td>
</tr>
<tr>
<td>Da-lin Bridge</td>
<td>ARIMA (8,1,1)</td>
<td>Training</td>
<td>0.3817</td>
<td>0.2280</td>
<td>30.8787</td>
<td>0.1779</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Testing</td>
<td>0.1460</td>
<td>0.3501</td>
<td>74.4639</td>
<td>0.2798</td>
</tr>
<tr>
<td>Shui-yuan Bridge</td>
<td>ARIMA (7,1,1)</td>
<td>Training</td>
<td>0.4097</td>
<td>0.2348</td>
<td>33.3921</td>
<td>0.1865</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Testing</td>
<td>0.1410</td>
<td>0.5629</td>
<td>80.5037</td>
<td>0.4312</td>
</tr>
<tr>
<td>Huang-ju-pi-liao</td>
<td>ARIMA (5,1,1)</td>
<td>Training</td>
<td>0.4049</td>
<td>0.2707</td>
<td>35.3087</td>
<td>0.1898</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Testing</td>
<td>0.1506</td>
<td>0.6653</td>
<td>78.4656</td>
<td>0.4730</td>
</tr>
</tbody>
</table>
basis or trial-and-error method. The limitation of these selection processes is that they are very time consuming and unreliable. The determination of input variables is instead based on the Box-Jenkins approach of using the PACF chart that was applied in this study (BuHamar et al. 2005). The delayed variables obtained from Box-Jenkins analysis are the most important variables and are used as input nodes in the network. In the case of Ping-lin station, the differenced time series of BOD5 depends on seven lagged variables, i.e., ARIMA (7, 1, 1); therefore, the number of input nodes is set equal to seven and the ANN model is given in the following form:

$$\Delta y_t = f(\Delta y_{t-1}, \Delta y_{t-2}, \ldots, \Delta y_{t-7})$$

After determining the input variables, ANN models with different number of nodes in the hidden layer are tested to determine the optimal architecture of the network. In order to prevent overfitting, one hidden layer is predetermined and 3–45 nodes in the hidden layer are tested to identify the optimal architecture in this study. The RMSE values for the training and testing datasets are used as a criterion to evaluate the performances of the potential models with different nodes. The model with minimum RMSE value for the training set and significant increasing of RMSE value for the testing set is accepted as the optimum among potential models. The testing results at Ping-lin station are shown in Figure 5(a), and the optimal number of nodes in the hidden layer is equal to nine. The selections of activation functions for the neurons in the hidden and output layers are evaluated on the basis of RMSE value for the training dataset. The result shows that applying the tansig activation function for the neurons in the hidden layer and purelin activation function for the neuron in the output layer could

![Figure 5](https://iwaponline.com/hr/article-pdf/47/5/1069/368623/nh0471069.pdf)

**Figure 5** The performance of: (a) different number of nodes in the hidden layer for the training and testing datasets at Ping-lin station; and (b) different combinations of activation functions used in the ANN model at Ping-lin station.
result in better model performance than others (Figure 5(b)). The learning rate and momentum coefficient are fixed to 0.25 and 0.8, respectively, after trial-and-error processes. In the case of Ping-lin station, an ANN model with seven input variables, one hidden layer with nine nodes, and activation functions of \textit{tansig-purelin} could give the optimal result (Table 3). Compared to the identified ARIMA model at Ping-lin station, higher value R and smaller values of RMSE, MAPE, and MAE suggest a better fit of the selected ANN model to the BOD\textsubscript{5} dataset.

The identifications of ANN models at other sampling stations followed the same procedures described previously for Ping-lin station. The optimized architecture of the ANN model at each sampling station and its associated performance are shown in Table 3. Regarding the testing datasets, the ANN models at Kuo-lai and Bi-hu stations performed the best and the model at Huang-ju-pi-liao station performed the worst. It is clear from the table that the R values for both training and testing datasets at all stations obtained from identified ANN models are higher than those obtained from identified ARIMA models. The RMSE, MAPE, and MAE values for both training and testing datasets at all stations obtained from identified ANN models are also smaller than those obtained from identified ARIMA models.

Figure 6 displays the entire time series for the simulated and measured BOD\textsubscript{5} at all sampling stations. A close pattern of variation among the simulated and measured values of BOD\textsubscript{5} is evident suggesting for the better predictive capability of the ANNs than that of ARIMA. Each ANN model not only depicts the trend of the data, but also roughly describes the extreme values which cannot be done by the ARIMA model. Due to the complexity of aquatic systems and abundant anthropogenic activities, ANN models with a more complex architecture are needed to compensate the nonlinear relationship within the process variables and could produce better fit of the measured BOD\textsubscript{5} concentrations. Although the improvement of model performances of ANNs is apparent, the overall ability is still not good enough (the averaged R value for the testing dataset is only about 0.5) and a more sophisticated method to model BOD\textsubscript{5} concentrations is still desired.

### Adaptive neuro fuzzy inference system

For ANFIS identifications, the original time series dataset of BOD\textsubscript{5} measured at each station is first differenced, the same data pre-processing being implemented as during the development of ANN models. One of the most important steps in developing an ANFIS model is still the determination of the input variables, and the selection procedure is also based on the examination of a PACF chart. In the case of Ping-lin station, i.e., ARIMA (7,1,1), seven lagged variables are

### Table 3 | The basic structure and performance of optimized ANN model for each sampling station

<table>
<thead>
<tr>
<th>Station</th>
<th>Structure (input-node-output)</th>
<th>Input variables (Δyt, ρ)</th>
<th>Dataset</th>
<th>R</th>
<th>RMSE</th>
<th>MAPE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ping-lin</td>
<td>7-9-1</td>
<td>p = 1–7</td>
<td>Training</td>
<td>0.6740</td>
<td>0.2473</td>
<td>69.2710</td>
<td>0.1802</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Testing</td>
<td>0.5608</td>
<td>0.4015</td>
<td>76.5663</td>
<td>0.2599</td>
</tr>
<tr>
<td>Dai-yu-ku Creek</td>
<td>5-12-1</td>
<td>p = 1–5</td>
<td>Training</td>
<td>0.6697</td>
<td>0.2962</td>
<td>64.0387</td>
<td>0.2002</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Testing</td>
<td>0.5620</td>
<td>0.3590</td>
<td>81.5944</td>
<td>0.2163</td>
</tr>
<tr>
<td>Jin-gua-liau Creek</td>
<td>5-15-1</td>
<td>p = 1–5</td>
<td>Training</td>
<td>0.6347</td>
<td>0.4064</td>
<td>69.1898</td>
<td>0.3145</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Testing</td>
<td>0.4677</td>
<td>0.4379</td>
<td>105.1526</td>
<td>0.3756</td>
</tr>
<tr>
<td>Kuo-lai</td>
<td>6-18-1</td>
<td>p = 1–6</td>
<td>Training</td>
<td>0.6947</td>
<td>0.2480</td>
<td>40.7337</td>
<td>0.2038</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Testing</td>
<td>0.4878</td>
<td>0.2605</td>
<td>55.1503</td>
<td>0.2071</td>
</tr>
<tr>
<td>Bi-hu</td>
<td>6-15-1</td>
<td>p = 1–6</td>
<td>Training</td>
<td>0.6579</td>
<td>0.2194</td>
<td>40.7533</td>
<td>0.1706</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Testing</td>
<td>0.5351</td>
<td>0.3035</td>
<td>56.6606</td>
<td>0.2187</td>
</tr>
<tr>
<td>Da-lin Bridge</td>
<td>8-9-1</td>
<td>p = 1–8</td>
<td>Training</td>
<td>0.6158</td>
<td>0.2037</td>
<td>25.8999</td>
<td>0.1516</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Testing</td>
<td>0.4968</td>
<td>0.3102</td>
<td>53.9803</td>
<td>0.2446</td>
</tr>
<tr>
<td>Shui-yuan Bridge</td>
<td>7-30-1</td>
<td>p = 1–7</td>
<td>Training</td>
<td>0.7833</td>
<td>0.2198</td>
<td>29.8330</td>
<td>0.1574</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Testing</td>
<td>0.4549</td>
<td>0.3829</td>
<td>46.0420</td>
<td>0.2874</td>
</tr>
<tr>
<td>Huang-ju-pi-liao</td>
<td>5-21-1</td>
<td>p = 1–5</td>
<td>Training</td>
<td>0.6670</td>
<td>0.2610</td>
<td>30.7647</td>
<td>0.1700</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Testing</td>
<td>0.3749</td>
<td>0.5458</td>
<td>56.6385</td>
<td>0.3990</td>
</tr>
</tbody>
</table>
selected as the inputs. After determining the appropriate input variables, ANFIS models with different number of rules are tested to determine the optimal architecture of system. The number of rules in ANFIS can be analogous to the number of nodes in ANNs; therefore, two to five rules are evaluated to identify the optimal architecture. The number of rules is increased one at a time to prevent overfitting. The RMSE values for the training and testing datasets are also used as a criterion to evaluate the performances of the potential models with different numbers of rules. Again, the model with minimum RMSE value for the training set and significant increasing of RMSE value for the testing set is accepted as the optimum among potential models. In the case of Ping-lin station, the result shows that the optimal number of rules is equal to three (Figure 7(a)). The selection from different MFs, i.e., Gaussian, bell-shaped, trapezoidal-shaped, and triangular-shaped functions, is also evaluated on the basis of the RMSE value for the training dataset. The result of Ping-lin station shows that applying bell-shaped MF could result in better model performance than others (Figure 7(b)). Therefore, in the case of Ping-lin station, an ANFIS model with seven input variables, three rules, and the bell-shaped MF could give the optimal results (Table 4). Compared to the identified ARIMA and ANN models at Ping-lin station, a considerable increase of R value and decreases of RMSE, MAPE, and MAE values indicate a significant improvement of predictive capability of the selected ANFIS model to the BOD$_5$ dataset.

The identification of ANFIS models at other sampling stations followed the same procedures described previously for Ping-lin station. The optimized architecture of ANFIS model at each of sampling station and its associated performance are shown in Table 4. Regarding the testing datasets, the ANFIS models at Da-lin Bridge station performed the best and the model at Huang-ju-pi-liao stations, again, performed the worst. It is clear from the table that the R values for both training and testing datasets at all
stations obtained from identified ANFIS models are higher than those obtained from identified ARIMA and ANN models. The RMSE, MAPE, and MAE values for both training and testing datasets at all stations obtained from identified ANFIS models are also smaller than those obtained from identified ARIMA and ANN models. Figure 8 displays the entire time series for the simulated and measured BOD$_5$ at all sampling stations. A very close pattern of variation among the simulated and measured values of BOD$_5$ is evident suggesting the good predictive capability of the selected model. Each ANFIS model not only depicts the trend of the data, but also describes the extreme values very well which cannot be done by either ARIMA or ANN model. These results indicate that ANFIS models are even more suitable than ANN models to compensate for the nonlinear relationship within the variables caused by the complexity of aquatic systems and to produce the best fit of the measured BOD$_5$ concentrations.

Comparisons of ARIMA/ANN/ANFIS models and other studies

In order to make this study more comprehensive, the capabilities of proposed ANFIS models are compared with ARIMA and ANN models in this section. The discrepancy ratio

<table>
<thead>
<tr>
<th>Station</th>
<th>Structure (input-rule-output)</th>
<th>Input variables ($\Delta y_p$)</th>
<th>Dataset</th>
<th>R</th>
<th>RMSE</th>
<th>MAPE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ping-lin</td>
<td>7-3-1</td>
<td>$p = 1-7$</td>
<td>Training</td>
<td>0.9345</td>
<td>0.2004</td>
<td>15.6456</td>
<td>0.0870</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Testing</td>
<td>0.8091</td>
<td>0.2247</td>
<td>32.5112</td>
<td>0.1165</td>
</tr>
<tr>
<td>Dai-yu-ku Creek</td>
<td>5-2-1</td>
<td>$p = 1-5$</td>
<td>Training</td>
<td>0.8829</td>
<td>0.2244</td>
<td>38.2991</td>
<td>0.1468</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Testing</td>
<td>0.8389</td>
<td>0.2427</td>
<td>41.1580</td>
<td>0.1564</td>
</tr>
<tr>
<td>Jin-gua-liau Creek</td>
<td>5-3-1</td>
<td>$p = 1-5$</td>
<td>Training</td>
<td>0.8255</td>
<td>0.3193</td>
<td>26.8040</td>
<td>0.1281</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Testing</td>
<td>0.8100</td>
<td>0.3397</td>
<td>42.8392</td>
<td>0.2096</td>
</tr>
<tr>
<td>Kuo-lai</td>
<td>6-2-1</td>
<td>$p = 1-6$</td>
<td>Training</td>
<td>0.9453</td>
<td>0.1124</td>
<td>20.5781</td>
<td>0.0689</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Testing</td>
<td>0.8935</td>
<td>0.1548</td>
<td>20.7423</td>
<td>0.0988</td>
</tr>
<tr>
<td>Bi-hu</td>
<td>6-3-1</td>
<td>$p = 1-6$</td>
<td>Training</td>
<td>0.9703</td>
<td>0.0971</td>
<td>14.3909</td>
<td>0.0496</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Testing</td>
<td>0.8381</td>
<td>0.1391</td>
<td>23.8338</td>
<td>0.1058</td>
</tr>
<tr>
<td>Da-lin Bridge</td>
<td>8-4-1</td>
<td>$p = 1-8$</td>
<td>Training</td>
<td>0.9712</td>
<td>0.0518</td>
<td>4.7980</td>
<td>0.0232</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Testing</td>
<td>0.8430</td>
<td>0.1215</td>
<td>10.2186</td>
<td>0.0627</td>
</tr>
<tr>
<td>Shui-yuan Bridge</td>
<td>7-2-1</td>
<td>$p = 1-7$</td>
<td>Training</td>
<td>0.9765</td>
<td>0.0628</td>
<td>3.1509</td>
<td>0.0189</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Testing</td>
<td>0.8487</td>
<td>0.1122</td>
<td>13.9100</td>
<td>0.0819</td>
</tr>
<tr>
<td>Huang-ju-pi-liao</td>
<td>5-3n-1</td>
<td>$p = 1-5$</td>
<td>Training</td>
<td>0.9662</td>
<td>0.1869</td>
<td>18.8125</td>
<td>0.0993</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Testing</td>
<td>0.5252</td>
<td>0.4083</td>
<td>47.0013</td>
<td>0.2441</td>
</tr>
</tbody>
</table>
$Dr = \frac{BOD_S}{BOD_m}$ is used for comparison of ARIMA, ANN and ANFIS results, where $BOD_S$ and $BOD_m$ are simulated and measured BOD$_5$, respectively. It is clearly seen from Figure 9 that more data points obtained from the ARIMA and ANN models are beyond the 50% error band (dashed lines), and it is evident that the ANFIS models can provide more accurate and reliable results than the ARIMA and ANN models. According to Figure 9, one can conclude again that applying ANFIS to model BOD$_5$ concentration improves the modeling accuracy compared to the ARIMA and ANNs.

Figure 8 | Simulated and measured BOD$_5$ of ANFIS models at: (a) Ping-lin; (b) Dai-yu-ku Creek; (c) Jin-gua-liau Creek; (d) Kuo-lai; (e) Bi-hu; (f) Da-lin Bridge; (g) Shui-yuan Bridge; and (h) Huang-ju-pi-iao.

Figure 9 | Distribution of the ARIMA, ANN, and ANFIS models discrepancy ratios. The data number 1 represents the first datum (January 1986) at Ping-lin station and the data number 2317 represents the last datum (May 2009) at Huang-ju-pi-iao station.
In the last 10 years, several studies have demonstrated a successful implementation of ANN modeling in the prediction of BOD concentrations, but very few cases employed ANFIS. The results obtained in this study are compared with those results reported in the ANN related literature. Oliveira-Esquerré et al. (2004) developed a multilayer perceptron neural network (MLP) combined with partial least square (PLS) method, called PLS-MLP, to predict inlet and outlet BOD of an aerated lagoon in Brazil. In their study, the PLS-MLP approach was the best of any method evaluated, and the results of BOD prediction in terms of coefficient of determination (R²) were 0.63–0.81. Singh et al. (2009) computed DO and BOD levels in the Gomti River, India using a three-layer feed-forward ANN model with the back propagation algorithm. The R² of simulated and measured BOD values for the training, validation, and testing sets were 0.85, 0.85, and 0.77, respectively. The respective values of RMSE for the three datasets were 2.25 for training, 1.84 for validation, and 1.38 for testing. Basant et al. (2010) described the ANN modeling for simultaneous prediction of the DO and BOD levels in the river water and the R² of simulated and measured BOD concentrations for the training, validation, and testing sets were 0.90, 0.76, and 0.74, respectively. Verma & Singh (2013) studied the ANN models to predict the BOD and COD concentrations and the corresponding R values were 0.976 and 0.981, respectively. In this study, the R values for the training and testing sets at all sampling stations were in the ranges 0.83–0.98 and 0.52–0.89, respectively, and the results are comparative to those reported in the previous literature. Compared to other countries worldwide, the variation of BOD₅ in Taiwan is relatively high and the excellent fitting of simulated results implies the powerful capability of the ANFIS method to the water quality prediction.

CONCLUSIONS

Water quality modeling of time series forecasting is an important, yet difficult, task that is required to make appropriate decisions about water systems. Despite the numerous time series models currently available, research for improvement on the effectiveness and accuracy of forecasting models has never ceased. This study demonstrates the potential of applying ANFIS in time series modeling of BOD₅ at the upstream catchment of Feitsui Reservoir, Taiwan for the first time. The systematic procedure of input selection and the pre-processing of input data during the model development are also presented.

In this paper, a large amount of BOD₅ samples was collected at eight sampling stations from 1986 to 2012 and developed into ANFIS models. The simulated results of ANFIS models were compared to those of both linear ARIMA and nonlinear ANN models. The data pre-processing of using differencing was implemented to the original time series data during the model development. The Box-Jenkins method of using a PACF chart to discover the appropriate lagged times in the ARIMA models was also applied and used to determine appropriate input variables in both ANN and ANFIS models. The performance of testing dataset was carefully examined during the training process of ANNs and ANFIS to avoid overfitting. Simulated results showed that the ANFIS model developed at each sampling station was superior to the ARIMA and ANN models in terms of the performances evaluated by R, RMSE, MAPE, and MAE. The excellent performance indicates the capability of the ANFIS method, particularly valuable for the relatively high variation of BOD₅ in Taiwan, and further stimulates the applications of the ANFIS method in other study areas.

ANFIS models can provide effective and accurate predictions for complex hydrological processes of BOD₅ because of its nonlinearity in nature and linguistic uncertainty. ANFIS can be seen as a predictive alternative to traditional modeling techniques and extended to other parameters of water quality or other areas to improve the understanding of river pollution trends. The time series modeling of using only time lagged variables can avoid the tedious work of collecting different types of water quality data, and the procedure of input selection and the pre-processing of input data presented in this study can intimate to other related studies. Further studies using more data from more watersheds may be required to strengthen these conclusions. Besides, the issue of climate change also has implications for the variations of BOD₅, and is a good candidate for future research.

ACKNOWLEDGEMENTS

This material is based on the work supported by National Science Council (NSC) under award NSC 102-2116-M-
019-001. The data supported by Taipei Feitsui Reservoir Administration and Taipei Water Management Committee are gratefully acknowledged.

REFERENCES


Maier, H. R., Jain, A., Dandy, G. C. & Sudheer, K. P. 2010 Method used for the development of neural networks for the


