Journal of Hydrology 388 (2010) 65–76

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

# Artificial neural networks for estimating regional arsenic concentrations in a blackfoot disease area in Taiwan

Fi-John Chang *, Li-shan Kao, Yi-Ming Kuo, Chen-Wuing Liu

*Department of Bioenvironmental Systems Engineering, National Taiwan University, Taipei 106, Taiwan, ROC*

ARTICLE INFO

*Article history:*
Received 17 August 2009
Received in revised form 12 April 2010
Accepted 14 April 2010

This manuscript was handled by L. Charlet, Editor-in-Chief, with the assistance of Eddy Y. Zeng, Associate Editor

*Keywords:*
Arsenic
Artificial neural networks (ANNs)
Over-fitting
Leave-one-out (LOO) cross-validation
Modified performance function (MPF)
Principal component analysis (PCA)

SUMMARY

High arsenic concentrations in groundwater have been detected in the south-western coastal area of Taiwan. In this study, artificial neural networks (ANNs) were investigated for their applicability to recovering the missing arsenic data and constructing the spatial distribution of arsenic concentration based on the arsenic concentration data of 28 groundwater observation wells. Due to a limited number of data sets, several strategies were proposed to construct the backpropagation neural networks (BPNs). The leave-one-out (LOO) cross-validation was adopted to diminish the bias in choosing validation data, and the modified performance function (MPF) was applied to reducing an over-fitting situation. Principal component analysis (PCA) was employed to transform the arsenic concentration of the regional wells into a limited number of main factors that were used as the input variables for the ANNs. Results showed that the LOO cross-validation was an effective tool for model selection, and the parameter, $\gamma$, of MPF played an important role for reducing errors in the model training and validation processes and alleviating the problem of over-fitting. Although sparse data sets have been used to construct ANNs, the models still achieved acceptable performance. The predicted spatial distribution of the arsenic concentration can provide useful information to local residents when groundwater achieves high levels of arsenic concentrations in non-functioning groundwater monitoring wells.



## 1. Introduction

Arsenic contamination in groundwater has led to a massive epidemic of arsenic poisoning in Bangladesh (Meharg, 2004), South Asia (Charlet and Polya, 2006.) and Taiwan (Chi and Blackwell, 1968). It is estimated that more than 137 million people in 70 countries are drinking arsenic-contaminated groundwater with the arsenic concentration value higher than the World Health Organization's standard. Smedley and Kinniburgh (2002) indicated that extreme arsenic concentrations in natural water are rare, but they are most frequently observed in groundwater, and the release from natural sources is the dominant cause of elevated arsenic concentrations in groundwater. Nath et al. (2009) stated the importance of hydrogeochemical characteristics of an aquifer in the release of arsenic to groundwater. Harvey et al. (2002) mentioned that young carbon has driven recent biogeochemical processes, and irrigation pumping is sufficient to have drawn water to the depth where dissolved arsenic is at a maximum. In Taiwan, groundwater is utilized abundantly as an alternative to surface water, including our study case – the Yun-Lin County where surface water resources are seriously deficient because of the high domestic, irrigational, aquacultural and industrial demands for water. Over-pumping introduces excess dissolved oxygen that may oxidize the immobile mineral, release arsenic and increase the arsenic concentration in groundwater (Liu et al., 2003). Therefore, the residents in the Yun-Lin County had used a high-arsenic artesian well for more than 50 years (Tseng, 1977), which exposed them to arsenic directly through drinking water or indirectly through various paths including ingesting aquacultural and agricultural products, thereby posing carcinogenic risks to human health (Liu et al., 2008). High arsenic concentrations in groundwater have also been verified to be associated with the blackfoot disease in Taiwan (Chiou et al., 1997). Although these hazards are genuine, a good management can reduce the risks. The Water Resources Agency installed 28 groundwater observation wells distributed in the coastal area of Yun-Lin County in 1992 to monitor the groundwater quality. The maintenance of these wells consumes massive labor and budget, and only four wells continued monitoring after 1999. A method that can be used to recover the missing data and estimate the data from the surrounding wells would be a useful tool to grasp the variations of groundwater quality.

Numerous methods for recovering missing data have been reported (Bennett et al., 1984; Hox, 1999; Chang et al., 2001; Little and Rubin, 2003). However, statistical methods such as linear regression, power and exponential methods are difficult to apply when trying to recover non-linear forms of the data set. Artificial

* Corresponding author. Tel.: +886 2 23639461; fax: +886 2 23635854.
*E-mail address:* changfi@ntu.edu.tw (F.-J. Chang).

0022-1694/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhydrol.2010.04.029

neural networks (ANNs) were inspired by biological neuron processing to perform brain-like computation through massively simple connected artificial neurons to identify the relationship between inputs and outputs of a system. Just as human beings apply knowledge gained from experience to new problems or situations, ANNs are capable of solving complex problems that might otherwise not have a tractable solution. The advantages of applying ANNs to water quality simulation are: (i) no physics-based algorithm is required to build the model; therefore, the modeling approach is faster and more flexible than physics-based modeling approaches in most cases; (ii) ANNs can handle non-linear relationship easily and properly; and iii) the expertise and user experiences may be incorporated easily into the model structure (Zhang and Stanley, 1997). Two critical issues in developing ANNs are: (1) how the model can be generalized to unseen data, and (2) how the model can be scaled with problem complexity. ANNs with more weights, such as having too many degrees of freedom in relation to the amount of data available, can cause over-fitting that will generally generate poor predictive performance, because it would usually amplify minor fluctuations in the data. To avoid over-fitting, it is necessary to use additional techniques (e.g. cross-validation and regularization) that can give an indication when further training is not resulting in better generalization.

There is a great concern about the potential effects of arsenic pollution on human health and environment; therefore, efficient modeling is crucial for accurate estimations of arsenic concentrations in the hydro-geological systems with limited data. In this study, ANN models were presented to recover the missing data at first, then to estimate the temporal and spatial variations of arsenic concentrations in the arsenic-polluted area. Groundwater quality data are usually monitored monthly or quarterly, which results in a small dataset. Relatively longer sampling intervals increase the difficulty in the prediction of groundwater quality using ANNs. High costs and low public awareness, however, are the main reasons for the lack of groundwater quality data in monitored fields. The effectiveness of ANNs could be significantly limited by the reduction of training data. We intend to implement a number of strategies – principal component analysis, cross-validation, and the modified performance function – to enhance the applicability and reliability of the constructed ANN models in the limited data circumstance. The PCA is usually used for transforming a large number of possibly correlated variables into a smaller number of uncorrelated variables, called principal components, to reduce the input dimensions effectively. The LOO cross-validation is adopted to diminish the bias in choosing training and validation data sets, and the MPF is applied to alleviating the over-fitting situation during the estimating process.

## 2. Study area

Yun-Lin County is located in the southwestern part of the alluvial fan of the Chou–Shui River (Fig. 1). The Chou–Shui River and Pei-Kong River are the two major rivers that flow through an area of approximately 1000 km$^2$ with extensions of 48 km from east to west and 24 km from north to south, respectively. Aquaculture and agriculture are the major incomes of local farmers in the coastal area of Yun-Lin County. Thus, a large amount of groundwater needs to be extracted from the aquifer to supply water to the fishponds and croplands. Over-pumping of groundwater for aquaculture leads to land subsidence, seawater intrusion, and soil salinization in this coastal area (Liu et al., 2001) so as to release arsenic

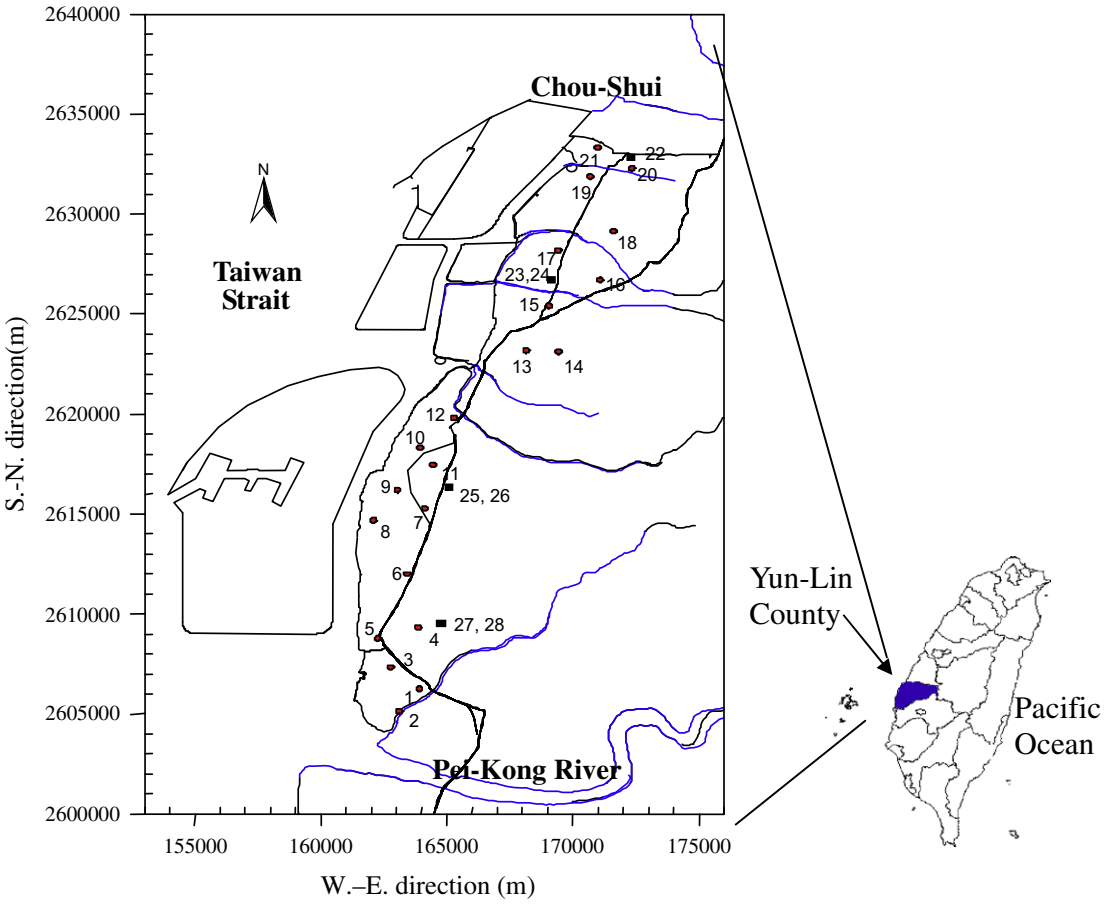


**Fig. 1.** Locations of the 28 groundwater wells in the Yun-Lin coastal area, Taiwan.

**Table 1**
The mean, standard deviation (SD), and coefficient of variation (*CV*) of arsenic concentration, well depth, and amount of data in the 28 groundwater wells.

| Well I.D. | No. of data[a] | No. of data[b] | Well depth (m) | Arsenic concentration (μg/L) | | | Well I.D. | No. of data[a] | No. of data[b] | Well Depth (m) | Arsenic concentration (μg/L) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Mean | SD | *CV* | | | | | Mean | SD | *CV* |
| #1 | 28 | 0 | 13.8 | 28.3 | 41.1 | 1.45 | #15 | 19 | 0 | 12.0 | 111 | 171.1 | 1.54 |
| #2 | 28 | 0 | 15.2 | 21.5 | 15.9 | 0.74 | #16 | 28 | 0 | 8.9 | 6.7 | 8.9 | 1.33 |
| #3 | 28 | 22 | 22.8 | 110 | 67.4 | 0.61 | #17 | 27 | 18 | 8.4 | 66 | 49.7 | 0.75 |
| #4 | 28 | 0 | 19.1 | 77.6 | 47.6 | 0.61 | #18 | 28 | 0 | 15.2 | 19.5 | 14.8 | 0.76 |
| #5 | 24 | 0 | 13.0 | 9.5 | 10.5 | 1.11 | #19 | 28 | 22 | 14.9 | 32.4 | 26.2 | 0.81 |
| #6 | 27 | 22 | 17.0 | 162 | 124 | 0.77 | #20 | 28 | 0 | 9.4 | 20 | 12.6 | 0.63 |
| #7 | 28 | 22 | 19.0 | 594 | 349 | 0.59 | #21 | 28 | 0 | 12.4 | 7.5 | 3.8 | 0.51 |
| #8 | 28 | 0 | 19.9 | 11.1 | 6.7 | 0.6 | #22 | 24 | 0 | 26.0 | 226 | 79.7 | 0.35 |
| #9 | 28 | 0 | 19.2 | 478 | 208 | 0.44 | #23 | 24 | 0 | 26.0 | 10.5 | 2.9 | 0.28 |
| #10 | 26 | 0 | 13.0 | 93 | 101 | 1.09 | #24 | 24 | 0 | 97.0 | 19.1 | 7.9 | 0.41 |
| #11 | 28 | 0 | 24.4 | 83.3 | 26.8 | 0.32 | #25 | 23 | 0 | 35.0 | 122 | 24.1 | 0.2 |
| #12 | 28 | 22 | 19.6 | 49.5 | 32.4 | 0.65 | #26 | 24 | 0 | 110 | 38 | 14.4 | 0.38 |
| #13 | 12 | 21 | 13.0 | 55.6 | 90.9 | 1.63 | #27 | 24 | 0 | 36.0 | 22.4 | 24.5 | 1.09 |
| #14 | 28 | 0 | 8.9 | 31.4 | 16.9 | 0.54 | #28 | 24 | 0 | 78.0 | 68.5 | 65.4 | 0.95 |

[a] The number of the data collected from 1992 to 1999.
[b] The number of the data collected from 1999 to 2005.

into groundwater (Liu et al., 2003). Hence, the Water Resources Agency constructed 28 groundwater wells in 1992, including 21 shallow wells and 7 deep wells, to monitor water quality variation in the coastal area. The depth of the wells ranged from 8 m to 110 m. A number of monitoring data from wells were obtained, and their relative information and results of fundamental statistical analysis are shown in Table 1. The wide ranges of mean, standard deviation and coefficient of variation (*CV*) of arsenic concentrations in the 28 wells indicate they are highly time-varying and spatially heterogeneous. Fig. 2 represents the variations of arsenic concentrations of the wells in three different regions. The central area of Yun-Lin County is arsenic-contaminated with higher concentration than the north and south areas. It is a tremendously difficult task to reconstruct the missing data based on these time-limited spatially heterogeneous data.

The arsenic concentration was determined by hydride generation followed by atomic absorption spectroscopy (APHA Method 3500-arsenic Part B). Groundwater samples were collected quarterly, and water quality data sets were obtained and analyzed (Tainan Hydraulic Laboratory, 1993–2005). Generally, four groundwater samples were collected each year from 1992 to 2005. Accordingly, from 1992 to 1999 we obtained arsenic data sets from 28 wells among which 15 decommissioned wells had complete data, while 13 decommissioned wells (#5, #6, #10, #13, #15, #17, #22, #23, #24, #25, #26, and #27) had incomplete data. The insufficient budget led to further shut-down of monitoring groundwater wells. From 1999 to 2005, only four wells (#3, #7, #12, and #19) remained functioning for groundwater sample collection. Some groundwater samples of the other 24 wells have been suspended for collection or lost (Table 1). Estimating the values of the lost groundwater samples is very important for realizing the variation of groundwater quality. The first step of this study is to estimate the missing data from 1992 to 1999 for these 13 wells by the constructed models, and then estimate the reliability of extended arsenic data of 24 wells from 1999 to 2005 based on the four wells (#3, #7, #12, and #19) with observed arsenic data from 1999 to 2005.

## 3. Methodology

### 3.1. Artificial neural networks, ANNs

ANNs are flexible modeling tools with the capability of learning the mathematical mapping between input and output variables of non-linear systems and generalizing the processes of control, classification and prediction. They are capable of providing a neurocomputing approach to solving complex problems. In the last decade, ANNs have been widely applied with success to various water resources problems, such as rainfall–runoff modeling (Antar et al., 2006; Chang et al., 2007; Chiang et al., 2007), flood control (Chang et al., 2008), ground water problems (Johnson and Rogers, 2000; Krishna et al., 2008; Nikolos et al., 2008), water quality (Chaves and Toshiharu, 2007; McNamara et al., 2008) and reservoir operation problems (Chang et al., 2005; Chaves and Chang, 2008). ANNs have recently been applied to recovering missing data in hydrology, meteorology, and water quality. Singh et al. (2004) applied a backpropagation neural network (BPN) to identifying the unknown pollution sources in groundwater with partially missing concentration observation data. He and Takase (2006) applied ANNs to estimating missing daily data for rainfall, flow discharge, and groundwater elevation. Diamantopoulou et al. (2007) applied the cascade correlation ANN models to estimating missing monthly values of water quality parameters in rivers. Coulibaly and Evora (2007) investigated various types of ANNs to fill in missing data of daily total precipitation records and daily extreme temperature series.

The BPN is a self-organizing, self-teaching and non-linear model which can be easily implemented in a wide variety of problems such as function approximation, time series forecasting, pattern recognition, and process control by using commercial computation software packages (e.g., MATLAB and R). BPNs have been applied in hydrology (ASCE Task Committee, 2000; Turan and Yurdusev, 2009; Teegavarapu and Chandramouli, 2005, Yang and Chang, 2005), meteorology (Ahmad and Simonovic, 2005; Venkatesan et al., 1997), and groundwater quality (Kuo et al., 2004; Almasri and Kaluarachchi, 2005; Yesilnacar et al., 2008).

The network topology used for the backpropagation algorithm is a fully connected, layered, and feedforward network. The network is divided into an input layer, a hidden layer, and an output layer. Each layer includes several neurons that are the fundamental building block for the network. The learning process goes from input layer to hidden layer and then to output layer by adjusting the connected weights and/or the non-linear transfer functions. The goal of learning is to determine a set of weights that will minimize the error function. The backpropagation algorithm for training the network is based on the steepest gradient descent method which computes the first derivative of a cost function with respect to the parameters (weights) of the network. The training process determines the BPN weights and is similar to the calibration of a

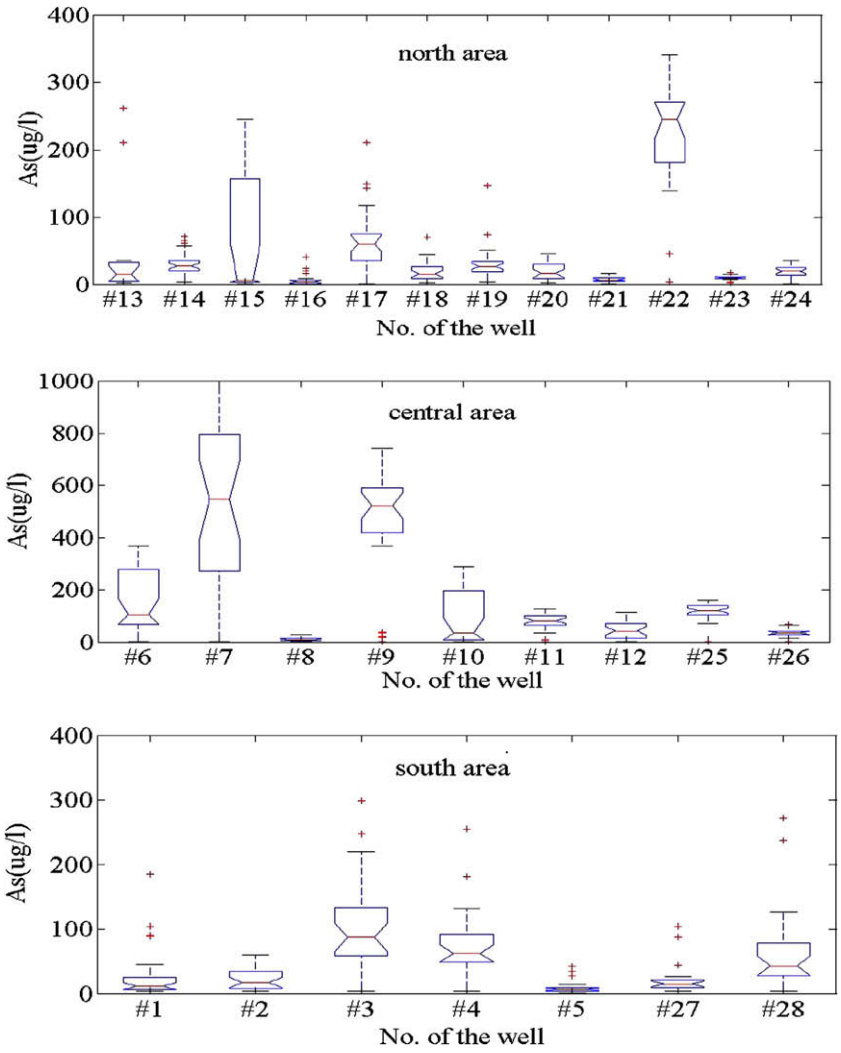


**Fig. 2.** The box-and-whisker plot of As concentration of the 28 wells in Yun-Lin coastal area. * A number of As concentrations in #7 and #15 wells are over range of the y-axis.

mathematical model. The BPNs are trained with a training set of input and known output data. At the beginning of training, the weights are initialized either with a set of random values or based on some previous experience. As training proceeds, the weights are systematically updated according to a training algorithm. The process is terminated when the difference between observed value and estimated value is less than a specified value and/or the number of iteration reaches a predefined number. Iteratively applying this method to adjusting the weights would gradually approach to a minimum of the cost function, but not necessarily the global minimum.

Neural networks are prone to over-fitting, especially when there are only a limited number of data. To avoid over-fitting, we implement a number of strategies – the leave-one-out cross-validation, the modified performance function, and the principal component analysis–to enhance the reliability of the constructed ANN models in the limited data circumstance.

### 3.2. The leave-one-out (LOO) cross-validation

Cross-validation is a direction towards estimating the generalization performance directly. Cross-validation, which consists of partitioning the data in training and test sets, is commonly used to obtain a reliable estimate of the test error for performance estimation or for use as a model selection criterion (Stone, 1974). It is the statistical practice of partitioning a set of sample data into subsets such that the analysis is initially performed on a single subset, while the other subset(s) are retained for subsequent use in confirming and validating the initial analysis. The LOO cross-validation is a special case of the famous *K*-fold cross-validation. In LOO cross-validation, each time one of the data is used as the validation set and the remaining data are put together to form a training set. The process is then repeated $n$ times. The LOO cross-validation has been shown to give an almost unbiased estimator of the generalization properties of statistical models, and therefore provides a sensible criterion for model selection and comparison (Adankon and Cheriet, 2009).

In this study, the LOO cross-validation was adopted to evaluate the predictive ability of the ANN models. Suppose that $n$ data sets are available for constructing ANNs and the data sets are randomly split into two groups, one for validation and the remaining $n-1$ for training. The $n$ subsets of the data sets are denoted $\{x_1, x_2, x_3, \cdots, x_n\}$. Each $\{x\}$ contains the same number of data. Each $\{x\}$ includes input data and output data. In the ANN model, input data are the known arsenic concentrations in the monitoring wells, and output data are the missing arsenic concentrations in the monitoring wells in the study area. The procedure of validation is to leave-one-out of the $\{x_1, x_2, x_3, \cdots, x_n\}$ as a validation data set, and the remaining $n-1$ subsamples are used as training data. The LOO cross-validation process is then repeated $n$ times with each

of the $n$ subsamples used exactly once as the validation data. The number of neurons in the hidden layer was changed to determine its effect on model validation. After validating the first well, we applied the same procedure to identifying and validating the ANN models of the other wells.

### 3.3. The modified performance function, MPF

When the ANN model is searching the optimal value of weights in the training process, over-fitting may occur in a limited input-output data condition. To improve the over-fitting in model selection, we propose to add a regularization term (Tikhonov and Arsenin, 1977; Li et al., 2007) to the model selection criterion which penalizes solutions where the networks' weights take on unduly large values. This commonly allows one to increase generalization capability in approximation problems (Tikhonov, 1963; Tikhonov and Arsenin,1977). The model selection criterion is then given by the traditional performance function, shown in Eq. (1), which must be modified to overcome the unsteady simulation results.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(o_i - d_i)^2 \tag{1}$$

where $n$ is the number of exemplars, $o_i$ is the network output value, and $d_i$ is the desired output value. The MPF, considered a regularization term in the model selection criterion (Tikhonov and Arsenin, 1977), is proven to efficiently reduce over-fitting when constructing the model (Solazzi and Uncini, 2004). The MPF combines errors between the desired output and the network output, and weights between the networks. The MPF is shown as follows:

$$MSE_{reg} = \gamma \cdot MSE + (1-\gamma)\cdot msw \tag{2}$$

where $\gamma$ is the performance ratio between network output error and weights: $msw = \frac{1}{m}\sum_{j=1}^{m} w_j^2$ is the square-sum of the values of the network weights, where m is the number of weights in the network. Using the MPF causes the network to have smaller weights and forces the network response to be smoother and less likely to over-fit.

### 3.4. Principal component analysis, PCA

PCA is commonly used as a tool in exploratory data analysis (Preisendorfer, 1988). It is a mathematical procedure to transform a number of correlated variables into a limited number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. We set arsenic concentrations of the monitoring wells as $X$, which is a D-dimensional random vector with covariance matrix $C$. The problem is to consecutively find the unit vectors $a_1, a_2, \cdots, a_D$ such that $Y_i = x^t a_i$. Suppose $(\lambda_i, u_i)$ are the pairs of eigenvalues and eigenvectors of $C$ such that $\lambda_1 \geqslant \lambda_2 \cdots \geqslant \lambda_D$ and $\|u_i\| = 1 \ \forall 1 \leqslant i \leqslant D$. Then $a_i = u_i$ and $\mathrm{var}(Y_i) = \lambda_i$ for $\forall 1 \leqslant i \leqslant D$. If the $x_1, x_2, \cdots x_n$ are given, the procedure of the method can be concluded as follows:

1. Compute $m = \frac{1}{n}\sum_{i=1}^{n} x_i$ by MLE.
2. Compute the covariance matrix $C = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)(x_i - m)^t$ by MLE.
3. Compute the eigenvalue/ eigenvector pairs $(\lambda_i, u_i)$ of $C$.
4. Compute the first $d$ principal components $u_i^j = x_i^t u_j$, for each observation $x_i$, $1 \leqslant i \leqslant n$ along the direction $u_i, i = 1, 2, \ldots, d$.

Following the procedure of the method, we analyzed the arsenic concentrations of the known wells in this area to form a few main variables called factors. And these factors were employed to be input information to the ANN model.

### 3.5. Evaluation of model performance

The performance of the simulation of training and validation sets is evaluated by following measures of goodness-of-fit: root mean squared error (*RMSE*) and Nash-Sutcliffe coefficient of efficiency ($C_{eff}$) shown in Eqs. (3) and (4), respectively. The Nash–Sutcliffe efficiency of 1 corresponds to a perfect match; an efficiency of zero indicates the model predictions are only as good as the mean of the observed data, whereas an efficiency less than zero means the average value of the observed data is a better predictor than that of the constructed model. The coefficient of efficiency ($C_{eff}$) can be used to assess the predictive power of constructed models. In addition, according to the EPA (Environmental Protection Administration) of Taiwan the safety standard for arsenic concentration of drinking water is 10 μg/l. The safety standard for arsenic concentration for usage in agriculture and aquaculture is 50 μg/l. These safety standards are also set as thresholds to evaluate the ANN model.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(o_i - d_i)^2} \tag{3}$$

$$C_{eff} = 1 - \frac{\sum(o_i - d_i)^2}{\sum(d_i - \mathrm{averge}(d_i))^2} \tag{4}$$

where $N$ is the number of exemplars, $o_i$ is the network output value, and $d_i$ is the desired output value.

## 4. Results and discussion

As mentioned above, we have implemented several methods for reconstructing and/or estimating the regional arsenic concentrations solely based on information of arsenic concentrations in nearby monitoring wells. We first investigated the correlation in terms of both space and time. The results show that arsenic concentrations of several wells of interest are not significantly related to arsenic concentrations of the neighboring wells. It indicates that the relationship of the spatially distributed arsenic concentration is complex and non-linear, and a linear model for estimating the variation of regional arsenic concentrations would be ineffective. Therefore, we employed ANNs to estimate the spatial variation of arsenic concentrations.

Because the missing data of the investigated wells occurred in different periods during 1992–2005, we have proposed two scenarios (cases) to reconstruct the arsenic concentration by ANNs. In Case I, we reconstruct the missing arsenic data in the 13 monitoring wells during the period of 1992–1999 by building ANN models with inputs of 2–5 factors computed by PCA. In Case II, the arsenic data of the 24 decommissioned monitoring wells from years 1999 to 2005 were rebuilt by using information of four nearby monitoring wells or two major factors computed by PCA.

In considering limited training data, we adopted LOO cross-validation to learn the reliability of the estimators in their performance and the MPF to relieve the over-fitting problem in the constructed ANN models. Another purpose of Case I is to make a comparison between the models configured with the MPF and without the MPF.

### 4.1. Estimation of missing arsenic data (Case I)

The mean, standard deviation (SD), and coefficient of variation (*CV*) of the arsenic concentration together with well depth and amount of data in the 28 groundwater wells are shown in Table

1. Among the 28 wells mentioned above, only 15 contained complete time-series data sets from 1992 to 1999. These 15 wells were used to estimate the missing data of the remaining 13 wells (#5, #6, #10, #13, #15, #17, #22, #23, #24, #25, #26, #27, and #28). But the ANN model with 15 input nodes would lead to an unduly large number of parameters (connected weights) for the model. Therefore, the PCA was first employed to transform original data to a few independent factors that are considered as the inputs for the ANN model.

We employed the PCA to transform the original data to a few independent factors. The arsenic concentrations of these 15 retained wells from 1992 to 1999 were analyzed by the PCA. Table 2 presents the eigenvalues and the percentages of variance associated with each factor. The first five main factors that would cumulate 76% of variance are considered as useful inputs. This also reveals that the first two factors explain approximately 47.4% of the total variance. Table 3a shows the loading of the six main factors for the model. Each loading of factors 1–4 concentrated on the variation of the arsenic concentrations associated with Wells #7, #9, #3 and #4, respectively. The loading of factor 6 did not largely concentrate on the variation of arsenic concentration of any specific well. Based on the above analysis, we constructed the estimation model with 1–5 main factors.

These main factors were used to construct 13 specific ANN models for the 13 wells with incomplete data based on their respective training periods (Fig. 3). In constructing the ANN model for the well with missing data, the ANN model has three parameters including the value of $\gamma$, the number of the hidden nodes, and the number of inputs, which need to be specified. We applied the LOO cross-validation with trial and error procedure to determining the optimal structure of the ANN for each well based on their *RMSE*s. In the LOO cross-validation process, if the well possessed "*n*" measured data, there would be *n* optimal ANNs for different training and validation data sets. The constructed ANNs were chosen according to the rule of the least *RMSE* of the validation sets in the process. The optimal value of $\gamma$ and the optimal structure and weights of the constructed ANN were combined to form an optimal ANN model through these procedures, which was then selected as the optimal ANN model for the investigated well. Finally, the optimized model for each well was used to estimate the missing data to complete the data sets in 1992–1999. Table 4 presents the number of data used to train and estimate the ANNs, the input factors, the number of the hidden nodes, the value of $\gamma$,and the corresponding results for each well. The optimal model in Table 4 is chosen in the LOO cross-validation process to estimate the missing arsenic data. In general, the results (i.e., *RMSE* and coefficient of efficient) are suitable in estimating arsenic concentrations of these wells.

To represent the general goodness-of-fit between the measured values with estimated arsenic concentrations, the variations of Wells #6 and #22, which have relatively high mean values and large standard deviations, are displayed in Fig. 4. It reflects that the ANN model can grasp the trend of the arsenic variation. However, the estimated values of a few seasons are away from the observed values in the Wells #6 and #22. We also found high *RMSE*s in the validation for Well #15 which only has 19 training data and relatively high coefficient of variation (*CV*) shown in Table 1. On the other hand, in other wells which have more training data and/or small *CV*, the constructed ANN models can provide acceptable results. Thus, the number of training data and variable variations significantly affect the reliability of ANN models. The results demonstrate that the LOO cross-validation method is an effective tool to choose a suitable network because it can assure that no val-

**Table 2**
Eigenvalues, percent of variance, cumulative eigenvalue, cumulative percent of variance for the factor analysis of arsenic concentrations of these wells in Yun-Lin coastal area, Taiwan.

| Factor | Eigenvalue | Percent of variance (%) | Cumulative of eigenvalue | Cumulative percent of variance (%) |
|---|---|---|---|---|
| 1 | 4.134 | 27.56 | 4.1 | 27.6 |
| 2 | 2.970 | 19.80 | 7.1 | 47.4 |
| 3 | 1.903 | 12.69 | 9.0 | 60.0 |
| 4 | 1.278 | 8.52 | 10.3 | 68.6 |
| 5 | 1.132 | 7.54 | 11.4 | 76.1 |
| 6 | 0.913 | 6.08 | 12.3 | 82.2 |
| 7 | 0.707 | 4.71 | 13.0 | 86.9 |
| 8 | 0.526 | 3.51 | 13.6 | 90.4 |
| 9 | 0.390 | 2.60 | 14.0 | 93.0 |
| 10 | 0.361 | 2.41 | 14.3 | 95.4 |
| 11 | 0.289 | 1.93 | 14.6 | 97.3 |
| 12 | 0.176 | 1.18 | 14.8 | 98.5 |
| 13 | 0.117 | 0.78 | 14.9 | 99.3 |
| 14 | 0.073 | 0.49 | 15.0 | 99.8 |
| 15 | 0.032 | 0.22 | 15.0 | 100.0 |

**Table 3a**
Scores for the six factors in model I (Case I).

| Well No. | Factor 1 | Factor 2 | Factor 3 | Factor 4 | Factor 5 | Factor 6 |
|---|---|---|---|---|---|---|
| #1 | 0.015 | −0.161 | 0.173 | 0.129 | −0.038 | −0.433 |
| #2 | 0.011 | −0.027 | 0.102 | −0.117 | −0.103 | −0.201 |
| #3 | 0.119 | −0.014 | **0.944** | −0.010 | −0.088 | 0.221 |
| #4 | 0.031 | −0.067 | −0.035 | **−0.941** | 0.043 | −0.026 |
| #7 | **0.983** | 0.128 | −0.119 | 0.026 | 0.000 | 0.012 |
| #8 | −0.004 | 0.001 | 0.032 | −0.027 | −0.017 | −0.106 |
| #9 | −0.127 | **0.970** | 0.043 | −0.060 | −0.044 | 0.004 |
| #11 | 0.024 | 0.070 | 0.050 | 0.121 | −0.002 | −0.519 |
| #12 | 0.015 | 0.064 | 0.138 | −0.022 | **0.790** | −0.398 |
| #14 | 0.021 | 0.027 | 0.133 | −0.078 | 0.072 | −0.107 |
| #16 | 0.011 | −0.004 | 0.006 | −0.057 | −0.169 | −0.083 |
| #18 | 0.010 | −0.023 | 0.029 | −0.182 | −0.054 | −0.095 |
| #19 | 0.010 | 0.035 | 0.014 | −0.052 | −0.558 | −0.504 |
| #20 | 0.012 | −0.013 | 0.097 | −0.134 | −0.031 | −0.042 |
| #21 | −0.003 | 0.002 | 0.012 | −0.012 | −0.051 | 0.001 |

**Table 3b**
Scores for the two factors in model II (Case II).

| Well No. | Factor 1 | Factor 2 |
|---|---|---|
| #3 | 0.146 | 0.981 |
| #7 | 0.989 | −0.148 |
| #12 | 0.025 | 0.125 |
| #19 | 0.001 | −0.028 |

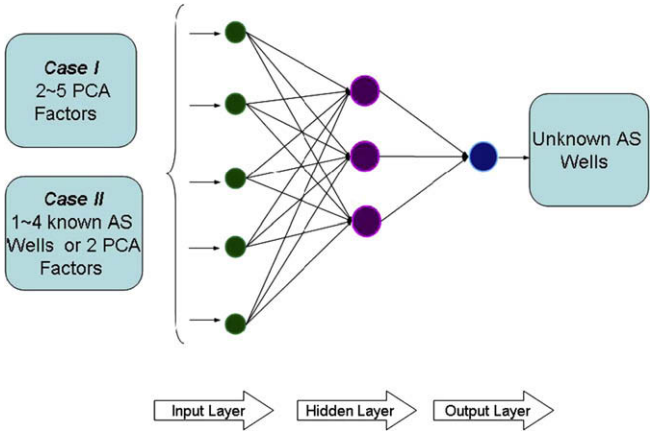


**Fig. 3.** The structures of BP neural networks that were used for recovering the missing data (Case I) from 1992 to 1999 and predicting the data (Case II) from 1999 to 2005.

**Table 4**
The structure and *RMSE* of the ANN for Case I in training and validation periods (unit: ug/l).

| Well I.D. | No. of the input Factors | Node | $\gamma$ | No. of data | | Average in the LOO process | | | | Optimal model | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | *RMSE* (μg/l) | | $C_{eff}$ | | *RMSE* (μg/l) | | $C_{eff}$ | |
| | | | | Observed | Estimated | Training | Validation | Training | Validation | Training | Validation | Training | Validation |
| #5 | 2 | 3 | 0.8 | 24 | 4 | 8.3 | 7 | 0.33 | −0.24 | 7.4 | 2.6 | 0.5 | 0.93 |
| #6 | 3 | 5 | 0.5 | 26 | 2 | 50.7 | 47.4 | 0.82 | 0.85 | 33.5 | 4 | 0.92 | 1 |
| #10 | 3 | 3 | 0.9 | 26 | 2 | 43 | 48.4 | 0.79 | 0.36 | 41.4 | 4.4 | 0.83 | 1 |
| #13 | 3 | 4 | 0.5 | 11 | 17 | 42.3 | 49 | 0.71 | −0.87 | 12 | 4.7 | 0.98 | 0.99 |
| #15 | 2 | 5 | 0.9 | 19 | 9 | 37.7 | 77.1 | 0.94 | −0.24 | 19.3 | 9.3 | 0.99 | 0.99 |
| #17 | 3 | 4 | 0.2 | 27 | 1 | 42.7 | 31.6 | 0.23 | 0.39 | 40.3 | 3.2 | 0.34 | 1 |
| #22 | 2 | 2 | 0.7 | 24 | 4 | 61.7 | 52.2 | 0.37 | 0.89 | 59.6 | 5.3 | 0.43 | 1 |
| #23 | 2 | 4 | 0.1 | 24 | 4 | 2.5 | 2 | 0.22 | 0.93 | 2.2 | 1.5 | 0.44 | 0.98 |
| #24 | 2 | 4 | 0.5 | 24 | 4 | 6.3 | 6.2 | 0.35 | 0.83 | 6 | 5.9 | 0.43 | 0.91 |
| #25 | 2 | 5 | 0.3 | 23 | 5 | 19.4 | 19.2 | 0.32 | 0.96 | 19.2 | 13.8 | 0.34 | 0.99 |
| #26 | 2 | 4 | 0.8 | 24 | 4 | 19.4 | 19.2 | 0.32 | 0.96 | 10.5 | 1.3 | 0.47 | 1 |
| #27 | 2 | 4 | 0.4 | 24 | 4 | 22 | 16.3 | 0.15 | −0.27 | 22.6 | 8.1 | 0.15 | 0.87 |
| #28 | 3 | 4 | 0.6 | 24 | 4 | 56.6 | 47.5 | 0.21 | −0.15 | 57 | 9.6 | 0.24 | 0.98 |

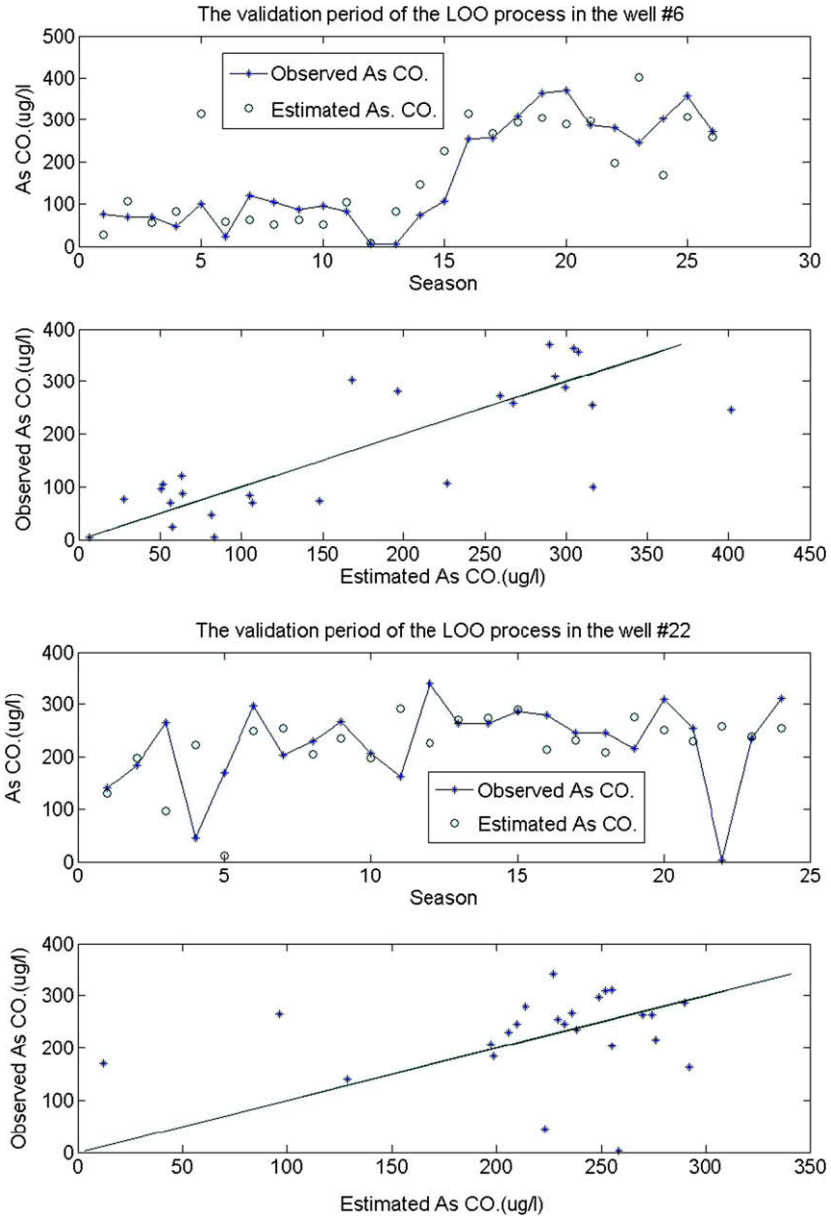


**Fig. 4.** The time series and scatter plot of the observed and estimated As concentration in the validation period of the LOO process in the wells #6 and #22.

idation data repeated in the training data set will provide more general information for the suitability of the constructed networks in validation patterns. Consequently, we suggest the LOO cross-validation should be applied to model construction processes, especially in the case of limited number of data for model construction.

To check the effect of the MPF which combines and balances the effects of errors and weights in the network construction stage, we constructed the same ANN models configuring with the MPF and without the MPF. The *RMSE*s of the two methods in the LOO cross-validation period are illustrated in Fig. 5. It appears if the MPF is adopted in the training period during model construction, the model performances are significantly improved by 18–78%. These results provide clear evidence that the MPF is able to mitigate the over-fitting problem, especially when a finite number of water samples are available in the high arsenic area (e.g., Wells #6, #15, and #22).

The effectiveness of the MPF is dependent upon the parameter $\gamma$, the performance ratio between output error and network weight (Eq. (2)). It is, however, difficult to determine the optimum value for the parameter. A larger value of $\gamma$ means a slight alleviation of the over-fitting situation in the training period of the model; while a smaller value of $\gamma$ means a strong alleviation. However, it can lead to a situation that the training data inadequately fits the network so as to cause a decrease in precision of the model in the predictive series. With the aim of evaluating its effort and obtaining the best performance function, the parameter is swept from 0.1 to 1. Fig. 6 shows the relationship between the $\gamma$ values with *RMSE* in validation periods selected from several monitoring wells where the node was adopted in ANN models according to Table 4. Thus, the arbitrary choice of $\gamma$ values may result in poor validations, especially in wells with high arsenic variations. In high arsenic concentration Wells #15 and #6, the optimal $\gamma$ values can be reduced by 20–30% of *RMSE*.

### 4.2. Estimation of arsenic data (Case II)

After 1999, many monitoring wells were suspended due to budget limitation; only four monitoring wells (#3, #7, #12, and #19) were retained and analyzed to form the two main factors shown in Table 3b by PCA. To estimate arsenic concentrations for the 24 decommissioned monitoring wells after 1999, the four monitoring wells were selected to construct the ANN models to estimate arsenic concentrations. If the model with the four known monitoring wells could not yield a suitable result, i.e., the estimation *RMSE* over the threshold (50 ug/l), we altered the input nodes from four monitoring wells to two major factors obtained from the PCA. We first constructed ANN models based on the four monitoring wells (or two main factors) for the 24 decommissioned wells with their respective results in training and validation period (1992–1999); the structure of the model is shown in Fig. 3 (Case II).

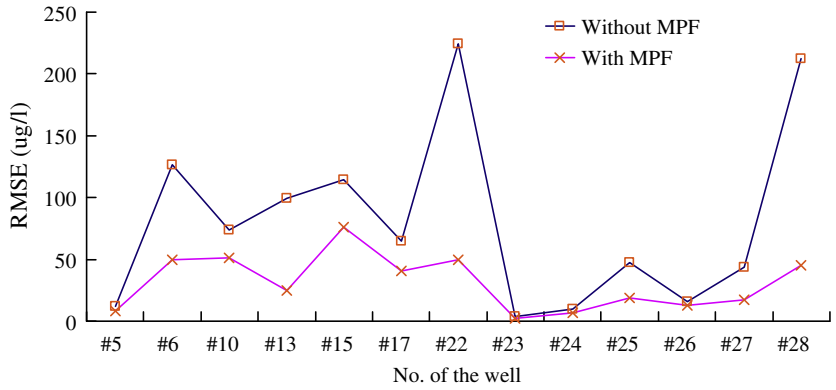


**Fig. 5.** The *RMSE*s in the validation period of the ANN models with/without applying the modified performance function.

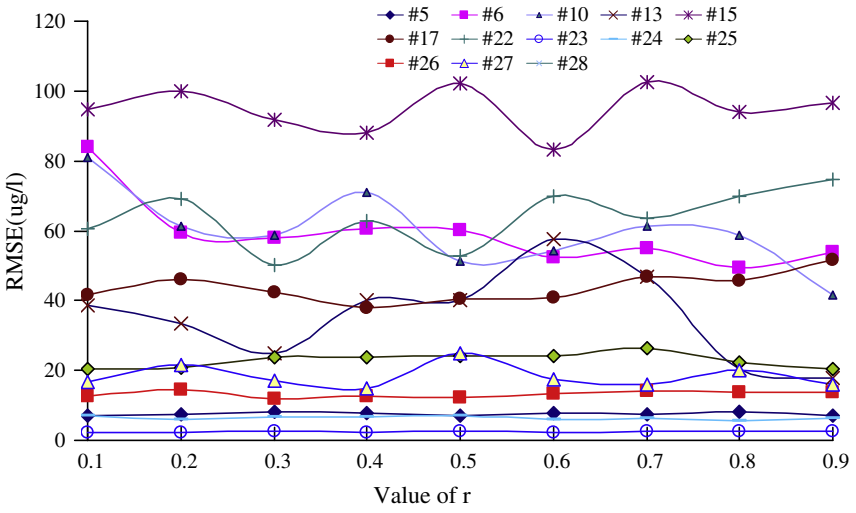


**Fig. 6.** The relationship between *RMSE* and $\gamma$ in validation periods for different wells.

According to the EPA of Taiwan, the safety standard for arsenic concentration for usage in agriculture and aquaculture is 50 μg/l. Thus, the safety standard for arsenic concentration (50 μg/l) was set to evaluate the performance of the ANN model. We considered this threshold (50 μg/l) in the training and validation periods to modify the input factors. If the *RMSE* of the model was too large (>50 μg/l), the parameters of the models (including input factors, the number of the nodes, and $\gamma$) would be altered. Therefore, the 24 constructed ANN models, which were selected from a large number of trial and error processes by using different input factors, a number of hidden nodes and $\gamma$ with the minimum *RMSE*s in the LOO cross-validation period, were then applied to estimating their arsenic variations.

The results of training and validation for 24 wells are shown in Table 5. In extending arsenic concentrations of the decommissioned monitoring wells, the arsenic concentrations of most unknown wells are estimated quarterly by using the four known wells. It appears that the four monitoring wells (#3, #7, #12, and #19) can reasonably figure out the variations of most wells. For example, the wells #5, #16, and #17 are located in low arsenic-contaminated areas where their validation *RMSE*s range from 7 to 28 μg/l within the error tolerance range; however, there are five wells (#6, #9, #10, #15, and #22) whose validation *RMSE*s are still greater than 50 μg/l (Table 5).

By inspecting these results, the missing data of most wells in general can be suitably estimated by the constructed ANNs, while the arsenic concentrations in wells #6, #9, #10, #15, and #22 can not be properly estimated. We notice that these wells have extremely high means and variations of arsenic concentrations. Another reason is the lack of relationship among wells in the area. We recognize that the arsenic concentration in a well is not only affected by the neighboring arsenic concentration but also by hydrological, geological and even biological environment variations. Nath et al. (2009) claimed the areas associated with high groundwater arsenic were associated with low Eh, and high Fe. Therefore, the large *RMSE*s of these wells could be imagined. In the circumstance, the ANN models could not provide reliable estimation for wells with high variations and short non-stationary time-series data. We note that if the sparse data with high variation cause the poor estimation, collecting sufficient data would be the most efficient and reliable method to solve the problem of poor estimation.

Comparing Table 4 with Table 5, the *RMSE*s of the poorly estimated wells (#10, #15 and #22) are lower in Case I (Table 4) than in Case II (Table 5). Besides, the arsenic concentrations in most wells (#5, #6, #10, #13, #15, #22, #24, #25, #27, and #28) are estimated more precisely in Case I (Table 4) than in Case II (Table 5). It indicates that sufficient information of arsenic concentrations of the neighboring wells can improve the accuracy of the model.

As we examine the Nash–Sutcliffe efficiencies shown in both Tables 4 and 5, we can easily tell that most of the coefficients of efficiency ($C_{eff}$) are greater than zero, which clearly indicates the predictive power of the constructed ANN models is valuable and effective. It, however, seems unavoidable that there are negative $C_{eff}$ obtained in a number of the wells under the validation processes. Closely checking the negative cases, we find they all have large coefficient of variation (*CV* greater than 0.95 shown in Table 1). In this circumstance, the ANN models cannot provide a reliable

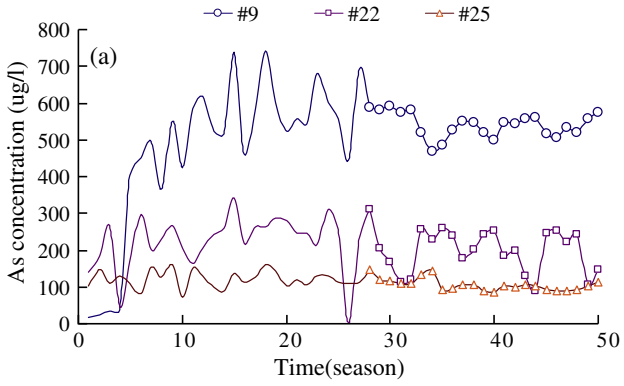


**Fig. 7.** The time-series As data of selected groundwater wells. Predicted values are illustrated after the 29 season for the wells.

**Table 5**
The structure and *RMSE* of ANN models that were used to predict the variations of arsenic concentrations for Case II in training and validation periods.

| Well I.D. | Input factor | Node | $\gamma$ | Average in the LOO process | | | | Optimal model | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | *RMSE* (μg/l) | | $C_{eff}$ | | *RMSE* (μg/l) | | $C_{eff}$ | |
| | | | | Training | Validation | Training | Validation | Training | Validation | Training | Validation |
| #1 | W3 | 4 | 0.2 | 38.8 | 27.8 | 0.07 | −1.32 | 39.3 | 2 | 0.09 | 1 |
| #2 | W3, W7, W12, W19 | 3 | 0.6 | 13.9 | 11.3 | 0.2 | 0.54 | 13.4 | 1.4 | 0.29 | 1 |
| #4 | W3, W7, W12, W19 | 3 | 0.5 | 41.2 | 35.7 | 0.21 | 0.52 | 33 | 31 | 0.51 | 0.84 |
| #5 | W3, W7, W12, W19 | 3 | 0.7 | 8.6 | 6.9 | 0.28 | −0.2 | 8.1 | 2.2 | 0.39 | 0.95 |
| #6 | W3, W7, W12, W19 | 7 | 0.6 | 71.5 | 62.6 | 0.56 | 0.81 | 59.5 | 43.6 | 0.7 | 0.94 |
| #8 | W3, W7, W12, W19 | 3 | 0.4 | 6.3 | 6.1 | 0.09 | 0.63 | 6 | 3.7 | 0.19 | 0.89 |
| #9 | W3, W7 | 4 | 0.9 | 136.6 | 146.6 | 0.54 | 0.83 | 118.3 | 43.9 | 0.68 | 0.99 |
| #10 | W3, W7, W12, W19 | 3 | 0.8 | 69 | 61.9 | 0.51 | 0.15 | 65.1 | 49.5 | 0.55 | 0.72 |
| #11 | W3, W7, W12, W19 | 3 | 0.7 | 22.9 | 22.5 | 0.23 | 0.88 | 21.3 | 35.7 | 0.34 | 0.82 |
| #13 | W3, W7, W12, W19 | 3 | 0.8 | 43.2 | 33.9 | 0.4 | −0.47 | 34 | 4.6 | 0.64 | 0.99 |
| #14 | W3, W7, W12, W19 | 3 | 0.1 | 14 | 11.5 | 0.29 | 0.77 | 15 | 1.1 | 0.21 | 1 |
| #15 | W3, W7, W12, W19 | 3 | 0.9 | 61.9 | 54.7 | 0.86 | 0.25 | 63.4 | 0.6 | 0.86 | 1 |
| #16 | W3, W7, W12, W19 | 3 | 0.1 | 7.8 | 5.4 | 0.2 | −0.71 | 7.6 | 3.3 | 0.26 | 0.75 |
| #17 | W3, W7, W12, W19 | 3 | 0.2 | 39.1 | 24.9 | 0.32 | 0.14 | 38.6 | 15.4 | 0.35 | 0.86 |
| #18 | W3, W7, W12, W19 | 3 | 0.6 | 13.7 | 11.7 | 0.11 | 0.33 | 38.6 | 15.4 | 0.35 | 0.86 |
| #20 | W3, W7, W12, W19 | 3 | 0.8 | 9.2 | 8.7 | 0.44 | 0.65 | 8.9 | 2.3 | 0.5 | 0.99 |
| #21 | W3, W7, W12, W19 | 3 | 0.2 | 3.5 | 3 | 0.12 | 0.77 | 3.2 | 3.1 | 0.2 | 0.83 |
| #22 | W3, W7, W12, W19 | 3 | 0.4 | 71.7 | 59.1 | 0.15 | 0.87 | 67.2 | 21.4 | 0.29 | 0.99 |
| #23 | W3, W7, W12, W19 | 3 | 0.3 | 2.2 | 1.9 | 0.42 | 0.94 | 2.5 | 1.5 | 0.21 | 0.98 |
| #24 | W3, W7, W12, W19 | 3 | 0.7 | 6.4 | 8.4 | 0.3 | 0.72 | 6.1 | 3.8 | 0.41 | 0.96 |
| #25 | W3, W7, W12, W19 | 3 | 0.9 | 19.6 | 21.9 | 0.31 | 0.95 | 18 | 21.7 | 0.44 | 0.97 |
| #26 | W3, W7, W12, W19 | 3 | 0.9 | 10.9 | 15.9 | 0.39 | 0.77 | 11.1 | 3.1 | 0.41 | 0.99 |
| #27 | W3, W7, W12, W19 | 3 | 0.6 | 22 | 15.6 | 0.15 | −0.14 | 21.8 | 6.5 | 0.21 | 0.91 |
| #28 | W3, W7, W12, W19 | 3 | 0.2 | 62 | 43 | 0.06 | 0.09 | 62.7 | 9.2 | 0.08 | 0.98 |

estimation for these wells with high variations and short non-stationary time-series data.

It is interesting to mention that in Tables 4 and 5 we also present the results (*RMSE* and $C_{eff}$) in both of the "Average" in the LOO cross-validation process and the optimal model. The optimal model is identified by selecting the best model (the least *RMSE*) from all of the *n* constructed models in the LOO cross-validation process. As expected, the optimal model has much better performance than the "Average" in the LOO cross-validation process. These results represent a situation one commonly faces when constructing the ANN through using different data sets for training and validation processes. We would like to note that the optimal model will only

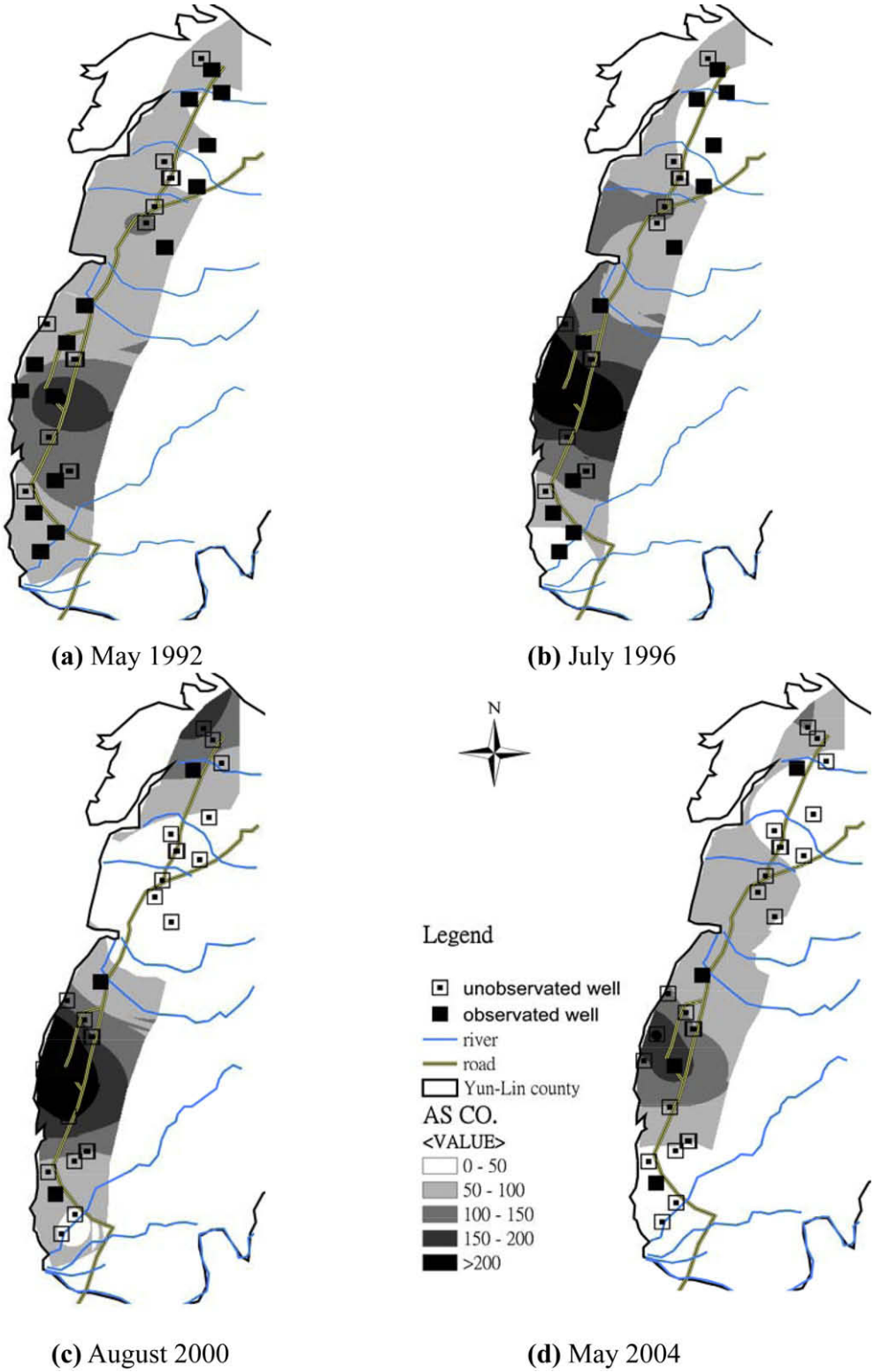


**Fig. 8.** The spatial distributions of As in four different years in the coastal area of Yun-Lin County.

be obtained by chance, while the LOO cross-validation process will give an unbiased estimator of the generalization properties of the models.

The estimated arsenic concentrations after 28 seasons of three selected wells are shown in Fig. 7. It illustrates three wells (#9, #22, and #25) containing high arsenic concentrations. These three wells seem to have the same patterns, where the peak concentrations were found in the 12th, 15th, and 18th seasons, and the low concentrations were found in the 10th, 14th, and 16th seasons. Based on the observed data with results obtained from this study, the spatial arsenic concentrations in the coastal area of the Yun-Lin County for 1992, 1996, 2000, and 2004 are shown in Fig. 8. The arsenic-polluted areas (defined by arsenic concentration >50 μg/l) were reduced from 1992 to 2004. In addition, the higher arsenic-polluted areas (defined by arsenic concentration >150 μg/l) near wells #6, #7, #9, and #11 were also narrowed from 1996 to 2004 (i.e., the contaminated domain was decreased). However, the highest concentration and contaminated domain was found near well # 21 (northern part of study area) in year 2000. This spatial distribution of arsenic concentration provides a warning to local residents when there are high levels of arsenic concentrations in the groundwater.

## 5. Conclusions

To alleviate the over-fitting problem and enhance the effectiveness of the constructed ANNs based on sparse groundwater data sets, we propose three strategies: (1) PCA for reducing the input dimensions, (2) LOO cross-validation for unbiased network selection, and (3) MPF for parsimonious network selection, by two scenarios (Case I and II) to reconstruct the missing arsenic concentrations. In Case I, we reconstructed the missing arsenic data of 13 wells using the regional information of 15 monitoring wells during the period of 1992–1999 by building ANN models with inputs of 2–5 major factors (from PCA). In Case II, we rebuilt (extended) the arsenic data of 24 decommissioned monitoring wells in the period of 1999–2005 by using information of four nearby monitoring wells or the two major factors by PCA. The LOO cross-validation proves to be an effective means of model selection for a variety of networks and has a higher probability to gain a reliable optimal model. The use of an MPF is shown to be effective in reducing over-fitting, and the number of weights of the network remains balanced and small, given a less complex network. The arsenic concentrations in most wells are estimated more precisely in Case I (Table 4) than in Case II (Table 5), which indicates that sufficient information of arsenic concentrations of the neighboring wells can improve the accuracy of the model.

The strategies used in this study for constructing the ANNs are useful and the problem of over-fitting is effectively alleviated. The ANN models achieve acceptable performance for most of the wells; however they cannot provide reliable estimation for those wells with high variations and short non-stationary time-series data. We demonstrate that the applicability and reliability of the ANNs are increased noticeably.

## Acknowledgments

This study is funded by the National Science Council, Taiwan, ROC, under Contract No. NSC 95-2313-B-002-051-MY3. The gauge data provided by the Water Resources Agency, Taiwan, are very much appreciated. The authors are grateful to the editors and two anonymous reviewers for their valuable comments and suggestions.

## References

Adankon, M.M., Cheriet, M., 2009. Model selection for the LS-SVM. Application to handwriting recognition. Pattern Recognition 42 (12), 3264–3270.
Ahmad, S., Simonovic, S.P., 2005. An artificial neural network model for generating hydrograph from hydro-meteorological parameters. Journal of Hydrology 315 (1–4), 236–251.
Almasri, M.N., Kaluarachchi, J.J., 2005. Modular neural networks to predict the nitrate distribution in ground water using the on-ground nitrogen loading and recharge data. Environmental Modelleing and Software 20 (7), 851–871.
Antar, M.A., Elassiouti, I., Allam, M.N., 2006. Rainfall-runoff modelling using artificial neural networks technique: a Blue Nile catchment case study. Hydrological Processes 20, 1201–1216.
Task Committee, ASCE, 2000. Artificial neural networks in hydrology II: hydrologic applications. Journal of Hydrologic Engineering 5 (2), 124–137.
Bennett, R.J., Haining, R.P., Griffith, D.A., 1984. The problem of missing data on spatial surfaces. Annals of the Association of American Geographers 74 (1), 138–156.
Chang, F.J., Hu, H.F., Chen, Y.C., 2001. Counterpropagation fuzzy-neural network for streamflow reconstructing. Hydrological Processes 15, 219–232.
Chang, F.J., Chiang, Y.M., Chang, L.C., 2007. Multi-step-ahead neural networks for flood forecasting. Hydrological Sciences Journal 52 (1), 114–130.
Chang, F.J., Chang, K.Y., Chang, L.C., 2008. Counterpropagation fuzzy-neural network for city flood control system. Journal of Hydrology 358, 24–34.
Chang, Y.T., Chang, L.C., Chang, F.J., 2005. Intelligent control for modeling of real time reservoir operation: part II ANN with operating curves. Hydrological Processes 19, 1431–1444.
Chaves, P., Chang, F.J., 2008. Intelligent reservoir operation system based on evolving artificial neural networks. Advances in Water Resources 31, 926–936.
Chaves, P., Toshiharu, K., 2007. Conceptual fuzzy neural network model for water quality simulation. Hydrological Processes 21 (5), 634–646.
Chi, I.C., Blackwell, R.Q., 1968. A controlled retrospective study of blackfoot disease, an endemic peripheral gangrene disease in Taiwan. American Journal of Epidemiology 88, 7–24.
Chiang, Y.M., Hsu, K.L., Chang, F.J., Yang, H., Soroosh, S., 2007. Merging multiple precipitation sources for flash flood forecasting. Journal of Hydrology 340, 183–196.
Chiou, H.Y., Hsueh, Y.M., Hsieh, L.L., Hsu, L.I., Hsu, Y.H., Hsieh, F.I., Wei, M.L., Chen, H.C., Yang, H.T., Leu, L.C., 1997. Arsenic methylation capacity, body retention, and null genotypes of glutathione S-transferase M1 and T1 among current arsenic-exposed residents in Taiwan. Mutation Research-Reviews in Mutation Research 386 (3), 197–207.
Charlet, L., Polya, D., 2006. Arsenic in shallow, reducing groundwaters in Southern Asia: an environmental health disaster. Elements 2, 91–96.
Coulibaly, P., Evora, N.D., 2007. Comparison of neural network methods for infilling missing daily weather records. Journal of Hydrology 341 (1–2), 27–41.
Diamantopoulou, M.J., Antonopoulos, V.Z., Papamichail, D.M., 2007. Cascade correlation artificial neural networks for estimating missing monthly values of water quality parameters in rivers. Water Resources Management 21 (3), 649–662.
Harvey, C.F., Swartz, C.H., Badruzzaman, A.B.M., Keon-Blute, N.E., Yu, W., Ali, M.A., Jay, J., Beckie, R., Niedam, V., Brabander, D.J., Oates, P.M., Ashfaque, K.N., Islam, S., Hemond, H.F., Ahmed, M.F., 2002. Arsenic mobility and groundwater extraction in Bangladesh. Science 298, 1602–1606.
He, B., Takase, K., 2006. Application of the artificial neural network method to estimate the missing hydrologic data. Journal of the Japan Society of Hydrology & Water Resources 19 (4), 249–257.
Hox, J.J., 1999. A review of current software for handling missing data. Kwantitatieve Methoden 62, 123–138.
Johnson, V.M., Rogers, L., 2000. Accuracy of neural network approximators in simulation-optimization. Journal of Water Resources Planning and Management – ASCE 126 (2), 48–56.
Kuo, Y.M., Liu, C.W., Lin, K.H., 2004. Evaluation of the ability of an artificial neural network model to assess the variation of groundwater quality in an area of blackfoot disease in Taiwan. Water Research 38 (1), 148–158.
Krishna, B., Rao, Y.R.S., Vijaya, T., 2008. Modelling groundwater levels in an urban coastal aquifer using artificial neural networks. Hydrological Processes 22, 1180–1188.
Li, W., Lee, K.H., Leung, K.S., 2007.Large-scale RLSC learning without agony. In: ACM International Conference Proceeding Series: Proceedings of the 24th International Conference on Machine learning 227, pp. 529–536.
Little, R.J.A., Rubin, D.B., 2003. Statistical analysis with missing data. Technometrics 45 (4), 364–365.
Liu, C.W., Huang, Y.K., Hsueh, Y.M., Lin, K.H., Jang, C.S., Huang, L.P., 2008. Spatiotemporal distribution of arsineic species of oysters (*Crassostrea gigas*) in the coastal area of Southwestern Taiwan. Environmental Monitoring and Assessment 138 (1), 181–190.
Liu, C.W., Lin, K.H., Kuo, Y.M., 2003. Application of factor analysis in the assessment of groundwater quality in a blackfoot disease area in Taiwan. Science of the Total Environment 313 (1–3), 77–89.
Liu, C.W., Lin, W.S., Shang, C., Liu, S.H., 2001. The effect of clay dehydration on land subsidence in the Yun-Lin coastal area, Taiwan. Environmental Geology 40 (4), 518–527.
McNamara, J.P., Kane, D.L., Hobbie, J.E., Kling, G.W., 2008. Hydrologic and biogeochemical controls on the spatial and temporal patterns of nitrogen and

phosphorus in the Kuparuk River arctic, Alaska. Hydrological Processes 22, 3294–3309.
Meharg, A.A., 2004. Arsenic in rice-understanding a new disaster for South-East Asia. Trends in Plant Science 9 (9), 415–417.
Nath, B., Chakraborty, S., Charlet, L., Stüben, D., Chatterjee, D., 2009. Mobility of arsenic in the sub-surface environment: an integrated hydrogeochemical study and adsorption characteristics of the sandy aquifer materials. Journal of Hydrology 364, 236–248.
Nikolos, I.K., Stergiadi, M., Papadopoulou, M.P., Karatzas, G.P., 2008. Artificial neural networks as an alternative approach to groundwater numerical modelling and environmental design. Hydrological Processes 22, 3337–3348.
Preisendorfer, R., 1988. Principal Component Analysis in Meteorology and Oceanography. Elsevier, New York.
Singh, R.M., Datta, B., Jain, A., 2004. Identification of unknown groundwater pollution sources using artificial neural networks. Journal of Water Resources Planning and Management 130 (6), 506–514.
Smedley, P.L., Kinniburgh, D.G., 2002. A review of the source, behavior and distribution of arsenic in natural waters. Applied Geochemistry 17, 517–568.
Solazzi, M., Uncini, A., 2004. Regularising neural networks using flexible multivariate activation function. Neural Networks 17, 247–260.
Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society, Series B 36, 111–147.
Tainan Hydraulic Laboratory, 1993–2005. The Yun-Lin Offshore Industrial Infrastructure Complexes Planning, Development and Monitoring Report Part I, Groundwater Level and Quality Measurements, vol. 6. National Cheng-Kung University, Taiwan, ROC.
Tikhonov, A.N., 1963. Solution of incorrectly formulated problems and the regularization method. Soviet Mathematics 4, 1035–1038.
Tikhonov, A.N., Arsenin, V.A., 1977. Solutions of Ill-Posed Problems. Winston&Sons, Washington.
Teegavarapu, R.S.V., Chandramouli, V., 2005. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. Journal of Hydrology 312, 191–206.
Tseng, W.P., 1977. Effects and dose-response relationships of skin cancer and blackfoot disease with arsenic. Environmental Health Perspectives 19, 110–119.
Turan, M.E., Yurdusev, M.A., 2009. River flow estimation from upstream flow records by artificial intelligence methods. Journal of Hydrology 369, 71–77.
Venkatesan, C., Raskar, S.D., Tambe, S.S., Kulkarni, B.D., Keshavamurty, R.N., 1997. Prediction of all India summer monsoon rainfall using error-back-propagation neural networks. Meteorology and Atmospheric Physics 62 (3), 225–240.
Yang, H.C., Chang, F.J., 2005. Modelling the combined open channel flow by artificial neural network. Hydrological Processes 19, 3747–3762.
Yesilnacar, M.I., Sahinkaya, E., Naz, M., Ozkaya, B., 2008. Neural network prediction of nitrate in groundwater of Harran Plain Turkey. Environmental Geology 56, 19–25.
Zhang, Q., Stanley, S.J., 1997. Forecasting raw water quality parameters for the North Saskatchewan River by neural network modeling. Water Research 31 (9), 2340–2350.