WATER RESEARCH 45 (2011) 5535–5544

Available at www.sciencedirect.com

ScienceDirect

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

Prediction of contamination potential of groundwater arsenic in Cambodia, Laos, and Thailand using artificial neural network

Kyung Hwa Cho [a,1], *Suthipong Sthiannopkao* [b,1], *Yakov A. Pachepsky* [c], *Kyoung-Woong Kim* [d], *Joon Ha Kim* [d,e,*]

[a] *Department of Civil and Environmental Engineering, University of Massachusetts, Amherst, 130 Natural Resources Road, Amherst, MA 01003, USA*

[b] *Department of Environmental and Occupational Health, National Cheng Kung University, Tainan City, Taiwan, ROC*

[c] *USDA-ARS, Environmental Microbial & Food Safety Laboratory, 10300 Baltimore Ave., Beltsville, MD 20705, USA*

[d] *School of Environmental Science and Engineering, Gwangju Institute of Science and Technology (GIST), 261 Cheomdan-gwagiro, Buk-gu, Gwangju 500-712, Republic of Korea*

[e] *Sustainable Water Resource Technology Center (SWRTC) at GIST, 261 Cheomdan-gwagiro, Buk-gu, Gwangju 500-712, Republic of Korea*

ARTICLE INFO

Article history:
Received 10 February 2011
Received in revised form
29 May 2011
Accepted 8 August 2011
Available online 27 August 2011

Keywords:
Multiple linear regression
Principal component regression
Artificial neural network
Principal component-artificial neural network

ABSTRACT

The arsenic (As) contamination of groundwater has increasingly been recognized as a major global issue of concern. As groundwater resources are one of most important freshwater sources for water supplies in Southeast Asian countries, it is important to investigate the spatial distribution of As contamination and evaluate the health risk of As for these countries. The detection of As contamination in groundwater resources, however, can create a substantial labor and cost burden for Southeast Asian countries. Therefore, modeling approaches for As concentration using conventional on-site measurement data can be an alternative to quantify the As contamination. The objective of this study is to evaluate the predictive performance of four different models; specifically, multiple linear regression (MLR), principal component regression (PCR), artificial neural network (ANN), and the combination of principal components and an artificial neural network (PC-ANN) in the prediction of As concentration, and to provide assessment tools for Southeast Asian countries including Cambodia, Laos, and Thailand. The modeling results show that the prediction accuracy of PC-ANN (Nash–Sutcliffe model efficiency coefficients: 0.98 (traning step) and 0.71 (validation step)) is superior among the four different models. This finding can be explained by the fact that the PC-ANN not only solves the problem of collinearity of input variables, but also reflects the presence of high variability in observed As concentrations. We expect that the model developed in this work can be used to predict As concentrations using conventional water quality data obtained from on-site measurements, and can further provide reliable and predictive information for public health management policies.

© 2011 Elsevier Ltd. All rights reserved.

* *Corresponding author*. School of Environmental Science and Engineering, Gwangju Institute of Science and Technology (GIST), 261 Cheomdan-gwagiro, Buk-gu, Gwangju 500-712, Republic of Korea. Tel.: +82 62 970 3277; fax: +82 62 970 243.
E-mail address: joonkim@gist.ac.kr (J.H. Kim).

[1] Cho and Sthiannopkao contributed equally to this work and are listed in alphabetical order.

0043-1354/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.watres.2011.08.010

1. Introduction

Groundwater resources are the most important components of drinking water supplies for Southeast Asian countries (Berg et al., 2007). Arsenic (As) contamination of groundwater has become a major problem on a worldwide scale because As is a carcinogenic element, which is mostly present as an inorganic species in natural water systems (Bagla and Kaise, 1996; AWWA, 2001; Berg et al., 2006; Kocar and Fendorf, 2009; Sthiannopkao et al., 2010). Naturally occurring As-enriched groundwater has been observed in tube well/hand pump drinking water supplies in South and Southeast Asia, including Thailand, Vietnam, Loa PDR, Cambodia, Myanmar, Bangladesh, India, Nepal, and Pakistan (Berg et al., 2001a,b; Smedley and Kinniburgh, 2002; Sun et al., 2002; Polya et al., 2003, 2005; Stanger et al., 2005; Kohnhorst, 2005; Tetsuro et al., 2006; Chiew et al., 2009). It is estimated that about 200 million people are at risk of harmful toxic effects of As in these countries (Sun et al., 2002).

In Southeast Asian countries, monitoring strategies of As contamination need to be improved in order to quantify As concentrations in groundwater, which can then be used to provide further information to better assess and manage public health. The detection of As contamination of groundwater resources, however, is hampered by a substantial labor and cost burden for Southeast Asian countries; it requires sophisticated equipment, highly skilled technicians, and incurs a high maintenance cost. As such, dealing with local As-associated problems in Cambodia and Laos remains problematic, where facilities for determining arsenic concentrations as well as human resources and funding for analyzing arsenic are still very much lacking.

Therefore, modeling approaches for As concentrations using on-site measurement data can be an alternative to characterizing the As contamination potential, to provide predictive information for better public health management. Indeed, obtaining on-site data such as pH, redox potential (Eh), and salinity are less expensive and much easier to perform than to detect and measure As contamination via graphite atomic absorption spectrophotometry (AAS) or inductively coupled plasma mass spectroscopy (ICP-MS). It is thought that these simple measurements can be used to fill the gap incurred by the missing facilities, human resources, and funding for arsenic determination using sophisticated equipment.

The most popular modeling approaches currently in use are multiple linear regression (MLR) and principal component regression (PCR), both of which attempt to find the most appropriate predictive model by fitting a linear equation to multiple observed data (explanatory variables). However, MLR and PCR may not be successful due to their statistical assumptions, including absence of outliers, normality, and randomness, which can be easily violated (Gros, 1997). In particular, the existence of correlations among explanatory variables (referred to here as "collinearity") diminishes the statistical stability (or robustness) and may cause significantly high prediction errors (Mac Nally, 2002). Moreover, incorporation of too many redundant (or insignificant), or too few, explanatory variables into models is not practical; thus, a stepwise regression technique is often required to determine the optimal number of variables to use as explanatory variables (over- or underspecification) (Luan et al., 2008). Furthermore, even though the predictive power of MLR and PCR can be acceptable, the models often result in the poor accuracies when making predictions with new datasets in the validation step.

Therefore, more complex nonlinear models have been developed that achieve better predictions, and have subsequently been applied to water resources problems such as hydrological processes, water quality problems, and dam operations (Maier and Dandy, 1996; Wen and Lee, 1998; Lee et al., 2003; Riad et al., 2004; Sarangi and Bhattacharya, 2005; Tayfur et al., 2005; Holmberg et al., 2006; Kuo et al., 2008). These nonlinear models such as artificial neural networks (ANNs), however, are not only difficult to construct, but also often cause over-fit problems in predictions; furthermore, the straightforward interpretation of relationships between explanatory and dependent variables cannot be achieved by nonlinear models. A few researchers have applied ANN to the prediction of As in groundwater (Purkait et al., 2008; Chang et al., 2010). For example, Purkait et al. (2008) used conventional water quality parameters to predict As contamination in Eastern India in an attempt to select the best model among linear and nonlinear models. Chang et al. (2010) applied an ANN model to recover missing values in As concentration datasets in an area of Taiwan.

This study utilizes a comprehensive dataset from three different countries, having a wide range of As concentration levels and conventional water quality parameters. The objective of this study is to analyze and evaluate the predictive performance of MLR, PCR, ANN, and the sequential combination of principal component and ANN (PC-ANN) in the prediction of As concentrations in groundwater and thereby provide improved assessment tools for Southeast Asian countries. In addition, a sensitivity analysis is also applied to investigate the cause and effect relationship between input parameters and As concentrations.

2. Materials and methods

2.1. *Field and sampling sites*

Groundwater samples were taken in an attempt to investigate the As concentration and five in-situ parameters for three countries: Cambodia, Loas, and Thailand (see Fig. 1). Table 1 shows the information on the timing and study sites. In Cambodia, thirty groundwater samples were collected from six villages in Kandal Province in 2008 ((Prek Thom village, Kbal Kaoh commune, $n = 5$), (Phoum Thom village, Phoum Thom commune, $n = 5$), (Chounlork village, Korki commune, $n = 5$), (Tuol Tnort village, Koki commune, $n = 5$), (Doun Sor village, Koki commune, $n = 5$), (Poul Pear Ker village, Khom Day Eth commune, $n = 5$)). In 2010, 49 samples were taken from south of Phnom Penh. In Laos, a total of 62 tube well samples were collected in 2008 from households located in provinces of Champasack ($n = 27$), Attapeu ($n = 10$), Saravane ($n = 11$), Savannakhet ($n = 4$), Borikhamxay ($n = 7$), and

Vientiane ($n = 3$). Vientiane is located in the upstream of the Mekong River, followed by Borikhamxay, Savannakhet, Saravane, Champasack and Attapeu, respectively. The areas along the Mekong River located in middle and southern parts of Laos starting from Borikhamxay downward are floodplain areas. In Thailand, the concentrations of As in groundwater were examined in Tambon Ongphra ($n = 10$) in Suphan buri province in 2008. In the three countries, samples for total As were collected from tube wells by following this sequence: 1) pumping the tube well for several minutes, 2) washing a clean polyethylene bottle with the well water, and 3) taking water without filtering.

2.2. *Sample analysis and on-site measurement*

As concentrations in groundwater samples were measured by graphite atomic absorption spectrophotometry (GF-AAS; Perkin Elmer 5100 PC, with a detection limit of 0.5 $\mu g\ L^{-1}$). GF-AAS was calibrated by an external standard technique in the range of 0–100 $\mu g\ L^{-1}$. For a quality control, standard reference material (SRM) for natural water (National Institute of Standards & Technology NIST 1640) was used to assure the precision of the measurement. After every tenth sample during analysis, the SRM sample and calibration standards were analyzed to check the analysis accuracy. All samples were measured at least twice in order to assess the measurement reliability; samples were reanalyzed if the error either from the SRM or from the calibration standards exceeded 10% or the relative standard deviation of the measurement exceeded 5%. Dilution was made with 2% HNO_3 when the concentration of the sample was over the upper limit of the standard range (100 $\mu g\ L^{-1}$). During the sample collection in three countries, a series of in-situ measurements were conducted: pH, Eh, water temperature (W_t) (HORIBA d-54 meter), electrical conductivity (EC), and total dissolved solids (TDS) (ORION 3 STAR, Thermo Electron Corporation).

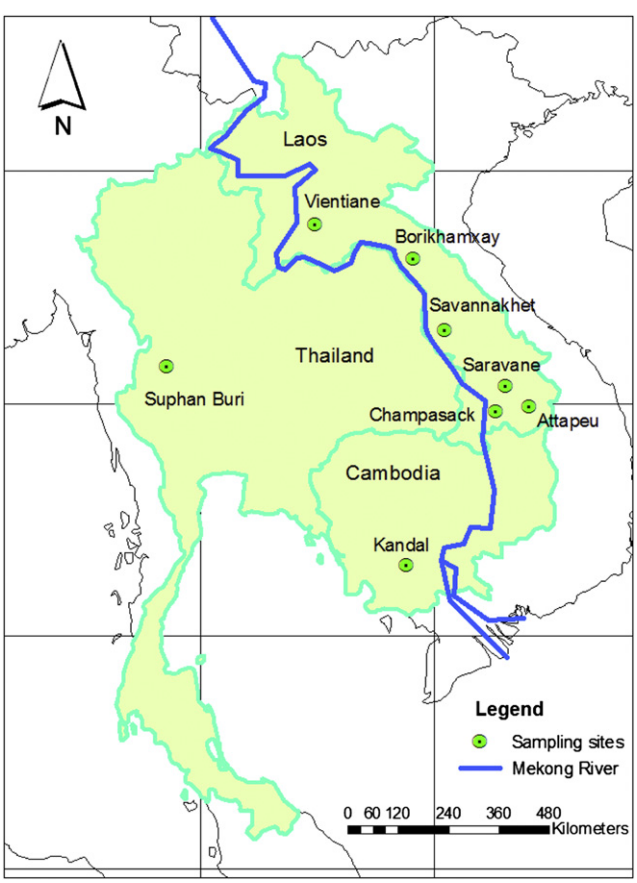

Fig. 1 – Map of study the site, showing sampling stations and Mekong River in Cambodia, Laos, and Thailand.

Table 1 – Timing and sites of Arsenic sampling in Southeast Asia.

Study area	Sampling time	Number of samples	Number of villages or provinces
Cambodia	2008	30	6
	2010	49	1
Laos	2008	62	6
Thailand	2008	10	1

2.3. *Modeling approaches: MLR, PCR, ANN, and PC-ANN*

All predictive models were developed using on-site data for pH, EC, TDS, W_t, and Eh. These variables were logarithmically transformed in order to normalize them for four different models (Rawlings et al., 1998). A pattern search algorithm and the Latin hypercube-one-factor-at-a-Time (LH-OAT) method were used to optimize parameters and investigate parameter sensitivity analysis of ANN, respectively (see Supplementary Information). Strategies of training, validation, and testing for ANN are also addressed in Supplementary Information.

Table 2 – Results of MLR and PCR for As concentration of groundwater.

Regression coefficients			Collinearity statistics	
Independent variable (i)	b_i[a]	Std. Error SE_{b_i}	VIF	Mean VIF
(Constant)[b]	2.83	2.73		2.02
pH	2.60	0.89	1.14	
Electrical conductivity	0.00	0.05	2.67	
Total dissolved solids	−0.23	0.04	2.83	
Temperature	−1.62	1.74	1.16	
Redox potential	−0.21	0.19	2.29	
(Constant)[b]	1.80	0.04		0.99
PC1	−0.18	0.04	0.98	
PC2	−0.13	0.04	0.99	
PC3	0.01	0.04	0.99	
PC4	−0.16	0.06	1.00	
PC5	−0.34	0.04	0.98	

a The subscript i indicates the water quality parameter, and b_i the computed coefficients of the water qualities in the MLR and PCR models,
$\log(As) = b_0 + b_1 \times pH + b_2 \times Conductivity + b_3 \times TDS + b_4 \times Water\ temperature + b_5 \times Redox$, $\log(As) = b_0 + b_1 \times PC1 + b_2 \times PC2 + b_3 \times PC3 + b_4 \times PC4 + b_5 \times PC5$.
b (Constant) in the table represents b_0 in the MLR and PCR models.

2.3.1. Multiple linear regression (MLR)

The MLR model was developed as follows:

$$\log(y) = \beta_0 + \sum_{i=1}^{n} \beta_i \log(x_i) \quad (1)$$

where x_i is the explanatory variable i, y is the dependent variable, β_i is the regression coefficient of explanatory variables i, and β_0 is the value of the intercept in the log-linear fitting.

2.3.2. Principal component regression (PCR)

PCR combines a principal component analysis (PCA) decomposition with MLR (Jolliffe, 2002; Cho et al., 2009). As a result of PCA, a new set of variables (the principal components (PCs)) and PC scores are generated from the orthogonal linear transform of the original data. The PC scores are then used in the regression as explanatory variables.

2.3.3. Artificial neural network (ANN)

ANN is a useful method for determining pattern classifications of multi-variable datasets as well as the prediction of complex processes. The multilayer perceptron ANN consists of two or more layers of nodes, including an input layer, a hidden layer, and an output layer, which are connected by links with varying weights. The nodal data are multiplied by the weights to compute the signal strength, and then are transferred to the next node in the network; the input layer nodes accept the input vectors and forward the signals to the next layer according to the connection. This process is continued until the signals reach the output layer.

2.3.4. Principal component-artificial neural network (PC-ANN)

PC-ANN merges PCA decomposition with ANN (Sousa et al., 2007). The main difference between this approach and ANN is that PC scores generated from the orthogonal linear transformation of the original data are used as the input variables of ANN; other procedures for the optimization, training, and validation are the same as those for the ANN model.

3. Results and discussion

3.1. Relationship between As concentrations and on-site measurements

Fig. 2 shows the average values and standard deviations of As concentrations and conventional water quality parameters in Cambodia, Laos, and Thailand. Arsenic concentrations of groundwater were observed to be the highest in Cambodia and the lowest in Laos. Similarly, pH was the highest in Cambodia, and the lowest in Laos. Conversely, Eh was the highest in Laos, and the lowest in Cambodia. Significant positive (Pearson correlation: 0.25; p-value: 0.00) and negative correlations (Pearson correlation: −0.10, p-value: 0.12) were found between the As concentrations and pH and between the As concentrations and Eh, respectively. Even though the p-value for Eh is greater than 0.05, the presence of this correlation is consistent with previous studies (Berg et al., 2007; Buschmann et al., 2007), indicating that high arsenic concentrations might be triggered by reducing conditions.

3.2. Linear models for predicting As concentrations

3.2.1. MLR for predicting As concentrations

Fig. 3(A) and (B) respectively shows the observed and the predicted As concentrations of groundwater in both the generation and validation steps of two different regression models (i.e., MLR and PCR). Overall, the MLR model did not reproduce the variations of observed As concentrations in either the generation or validation steps; the developed MLR model tends to underestimate As concentrations. Table 2 shows the regression coefficients b_i, the corresponding

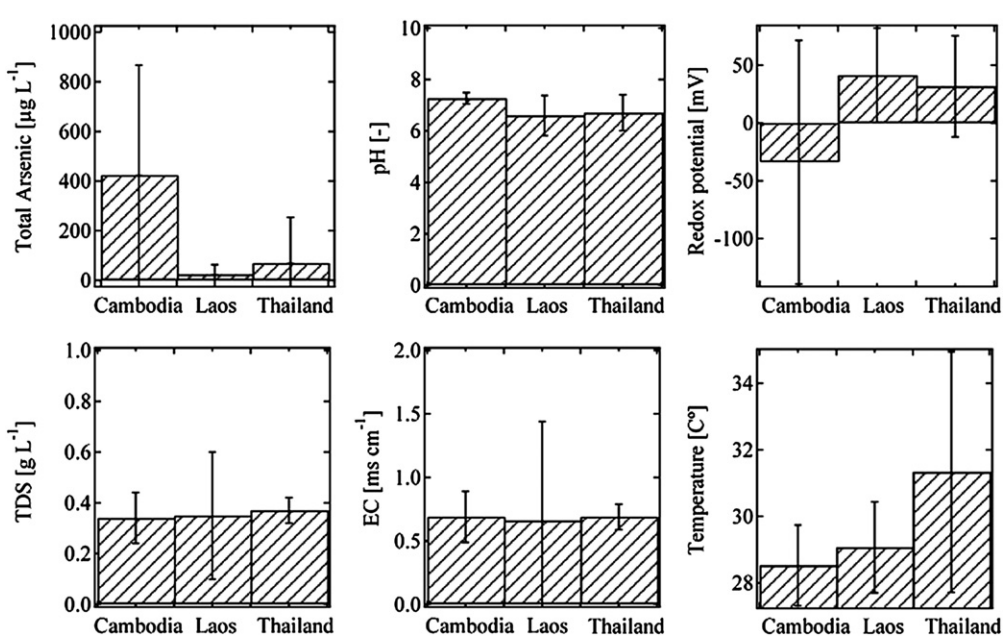

Fig. 2 – Mean values of As concentration, pH, redox potential, total dissolved solids, electrical conductivity, and temperature along with their standard deviations for Cambodia, Laos, and Thailand.

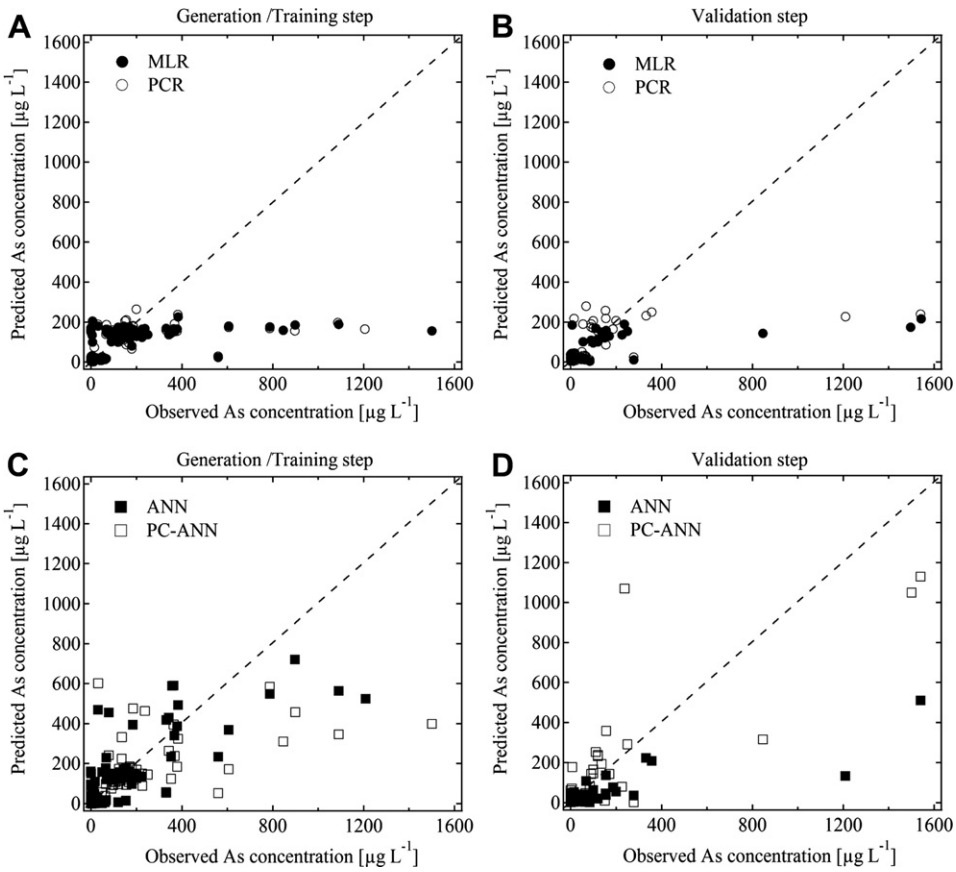

Fig. 3 – Comparison results between the observed and predicted As concentrations of groundwater, (A) and (B): MLR and PCR, (C) and (D): ANN and PC-ANN, closed circles: MLR, open circles: PCR, closed squares: ANN, open squares: PC-ANN.

standard error SE_{b_i}, and the collinearity statistics of the MLR model. Note that if the absolute value of b_i in the MLR model is greater than twice its standard error (i.e., SE_{b_i}), the ith variable can be regarded as a significant variable (Rawlings et al., 1998). Here, TDS and pH have regression coefficients greater than twice their standard errors.

Furthermore, collinearity was found among the explanatory variables because the largest variance inflation factor (VIF) was greater than 10, with an average VIF value being substantially greater than 1 (Bowerman and O'Connell, 1990; Myers, 1990). Consequently, the computed values of b_i and SE_{b_i} for a certain explanatory variable strongly rely on the degree of its correlation with the other variables in the MLR model. The modeling accuracy was also compared using the Nash–Sutcliffe model efficiency coefficient (NSE) computed from the predicted and observed As concentrations (Nash and Sutcliffe, 1970). In essence, the closer the NSE is to 1, the more accurate the model is. In particular, an acceptable NSE value needs to be greater than 0.5, while a good agreement value should be greater than 0.7 (Moriasi et al., 2007). As shown in Table 3, the NSE values of the MLR model (−0.29 and −0.75) indicate that the MLR model developed in this study cannot be considered a suitable model for predicting the As concentration of groundwater using on-site measurement data. This result is similar to a previous study in Eastern India; Purkait et al. (2008) found that MLR showed good accuracy for low levels, but that it did not reproduce the high levels of As concentration.

3.2.2. *PCR for predicting As concentrations*

In the PCR model, five PCs were generated from five water quality parameters and then used as input variables for

Table 3 – MAEs and Nash–Sutcliffe model efficiency coefficients of the model predictions for As.

		Generation/Training steps		Validation step	
		MAE[a]	NSE[b]	MAE[a]	NSE[b]
MLR		141.59	−0.29	345.26	−0.75
PCR		93.68	0.14	102.54	0.19
ANN		79.68	0.71	88.93	0.47
PC-ANN		72.08	0.61	101.17	0.66
Group A	ANN	24.93	0.83	24.78	0.34
	PC-ANN	22.63	0.84	44.09	0.52
Group B	ANN	49.68	0.96	111.21	0.74
	PC-ANN	35.10	0.98	75.15	0.71

a Mean absolute error (MAE) is a statistical approach used to assess the model performance, and its unit is μg L^{-1} $MAE = [n^{-1}\sum_{i=1}^{n}|x_{obs} - x_{pre}|]$ where n indicates the number of observations of As. Here, x_{obs} and x_{pre} indicate the observed and predicted As concentrations, respectively.

b $NSE = 1 - \frac{\sum_{t=1}^{T}(x_{obs}^{t} - x_{pre}^{t})^2}{\sum_{t=1}^{T}(x_{obs}^{t} - \overline{x}_{pre}^{t})^2}$ where x_{obs} is the observed As concentrations, x_{pre} is the predicted As concentrations, and $\overline{x}_{obs}^{t}$ is the averaged value of the observed As.

predicting As concentrations. Here, the cumulative percentage of the variations explained by the five PCs was 100%, implying that these PCs were able to explain all the variations in the original dataset. The five extracted PCs are significantly related to each water quality parameter, PC1 (Eh, 0.95), PC2 (conductivity, 0.93), PC3 (water temperature, 0.99), PC4 (pH, 1.00), and PC5 (TDS, 0.81).

Fig. 3(A) and (B) respectively compares the observed and the predicted As concentrations of groundwater in both the generation and validation steps of PCR. Overall, however, the PCR model did not reproduce variations of the observed As concentrations, usually by underestimating the relatively high As concentrations similar to the MLR. As shown in Table 2, PC1, PC2, PC4, and PC5 had regression coefficients with absolute values being greater than twice their standard errors. In addition, all VIF values for the explanatory variables were equal to or less than 1, thereby implying that the collinearity problem in MLR was completely overcome through the PC application. As shown in Table 2, however, the calculated NSEs imply that the PCR model developed in this study still cannot be used as a reasonable model for predicting groundwater As concentrations using on-site measurement data. Consequently, it demonstrates that the linearity of MLR and PCR cannot reproduce the dynamic variations of observed As concentrations in groundwater.

3.3. *Nonlinear models for predicting As concentrations*

3.3.1. *Optimization processes for ANN and PC-ANN models*

The optimal momentum rate, the number of hidden nodes, and the learning rate were obtained by the pattern search algorithm. The pattern search was used to determine the optimal parameter set from the ranges of three parameters; hidden nodes were ranged from 5 to 20, and learning and momentum rates were ranged from 0.01 to 0.6. Finally, the pattern search provided the optimal parameter set of the learning and momentum rates and the number of hidden layers, resulting in a minimum objective function. Moreover, while too few hidden nodes in ANN results in a poor predictive power, too many hidden nodes may cause a large computational time and over-fitting. The optimized number of hidden nodes for the ANN and PC-ANN models were respectively set to 13 and 15, which is consistent with past works on ANN because they range between 2 and 3 times the number of input nodes (Brion and Lingireddy, 1999). Note that the "tan-sigmoid" transfer function was used in the neurons of the hidden and output layers.

After the determining the learning rate, the momentum rate, and the number of hidden nodes in the ANN and PC-ANN models, the errors for the prediction of As concentrations were computed in terms of MAE and NSE, and then compared to those of MLR and PCR.

3.3.2. *ANN for predicting As concentrations*

Fig. 3(C) and (D) compares the observed and the predicted As concentrations of groundwater in both the training and validation steps of ANN, where the horizontal and vertical axes respectively indicate the observed and the predicted As concentrations in groundwater. Overall, it can be seen that the ANN model well reproduced the variations of the observed As concentrations of groundwater in Cambodia, Laos, and Thailand. As shown in Table 2, NSE values of ANN in training and validation step are respectively greater than 0.5 and approximately 0.5. In this case, it is clear that the MAE values for ANN are much less than those for MLR and PCR. Consequently, it can be seen that the optimized ANN model can be a useful tool in predicting groundwater As concentrations using on-site measurement data. This result is in good agreement with a previous study by Purkait et al. (2008), in which ANN showed better results than either a linear model.

3.3.3. *PC-ANN for predicting As concentrations*

Fig. 3(C) and (D) illustrates the observed and predicted groundwater As concentrations in both the training and validation steps of PC-ANN, where the open circles and squares indicate values from PC-ANN. The figure shows that the PC-ANN model also well reproduced the variations of observed groundwater As concentrations in Cambodia, Laos, and Thailand. As shown in Table 2, whereas the NSE value for PC-ANN in the training step was 0.61, the NSE in the validation step was 0.66; i.e., the NSE of PC-ANN was greater than that of ANN in the validation step. Consequently, it demonstrates that the PC-ANN model can be useful in predicting groundwater As concentrations using on-site measurement data.

3.4. *Comparison between MLR, PCR, ANN, and PC-ANN*

As shown in Table 2, the prediction accuracies of ANN and PC-ANN are better than those of MLR and PCR models in both the generation/training and validation steps. The table also demonstrates that the developed ANN and PC-ANN models show acceptable accuracies for predicting As concentrations in the groundwater in Cambodia, Laos, and Thailand, which can be explained by the fact that the nonlinearity of ANN can reproduce the vigorous variations of As concentrations. As mentioned above, the MLR and PCR models developed in this study are not deemed to be suitable models because linearity is not sufficient for explaining the dynamic variations of observed As concentrations. This study demonstrates, however, that the ANN model is a more acceptable approach than the MLR model in terms of modeling As concentrations using data from on-site measurements. In particular, it can be posited that PC-ANN is the best model because of its highest NSE in the validation step.

3.5. *Redox potential vs As concentration*

Fig. 4 shows the As concentration versus Eh. Based on the Eh, the dataset was divided into two groups, Group A (<0.00) and Group B (≥ 0.00). The 89.23% of data for Group A are samples collected from Kandal Province in Cambodia and the 84.52% of data for Group B were measured from Lao PDR and Thailand. Kandal Province is located in the Mekong Delta, which receives a substantial volume of sediment (160 millon t yr^{-1}) from the Mekong River (Meybeck and Carbonnel, 1975; Ta et al., 2001); the delta is mainly composed of young alluvial soil (Nguyen et al., 2000). In addition, a reduced state of As (III) was found to be the dominant species in Kandal Province (Polya et al., 2003; Rowland et al., 2008; Polizzotto et al., 2008). Consequently, Group A in the figure can be characterized as

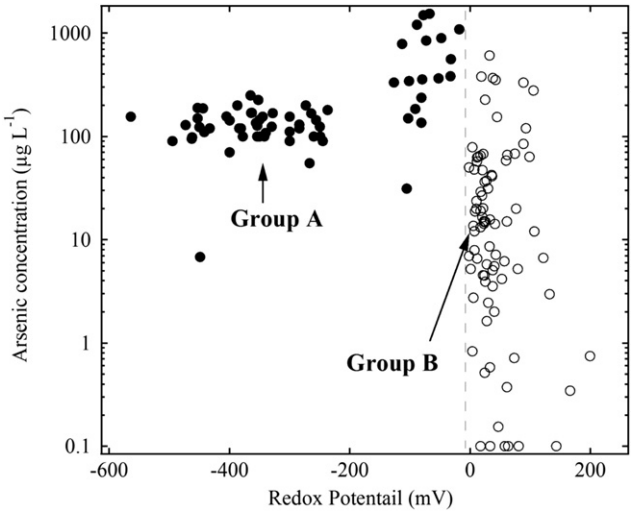

Fig. 4 – As concentration versus redox potential, showing Groups A and B.

predominantly As in highly reducing aquifer regions having a low Eh level. In the figure, the As concentrations are mostly stable in the elevated level (Group A), but show a steep negative trend with increasing Eh in Group B (i.e., Lao PDR and Thailand). Hydrological conditions in Lao PDR, Cambodia, and Vietnam have many characteristics in common and As (III) was also found to be a dominant species in Lao PDR. Here, Group B can also be interpreted as As in a reducing region, though it showed a negative correlation with Eh as opposed to Group A. This correlation may result from a combination of low iron (ferrous) concentrations and high Eh in Lao PDR (Chanpiwat et al., 2011). This different type of dependence of As on Eh found may result in difficulty in using the entire dataset for a single modeling approach. Therefore, Groups A and B were utilized to train ANN and PC-ANN models, which showed superior accuracy than linear models, in an attempt to improve the prediction performance.

Fig. 5 compares the observed and predicted values of As concentrations for the two datasets (Groups A and B). The two models show a better prediction for Group A, but relatively poor accuracies for Group B. Table 2 also demonstrates that the NSE of the models for Group A were lower than the models for Group B in the validation steps, and even worse than the

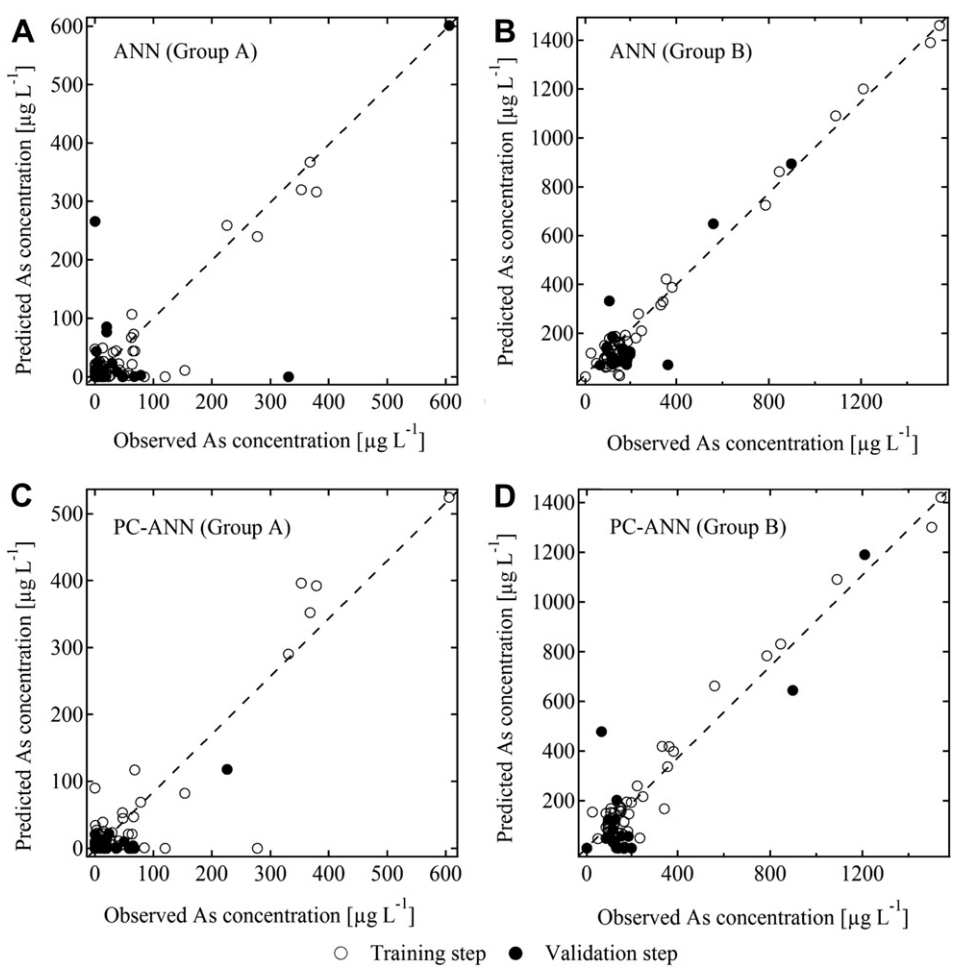

Fig. 5 – The prediction results of ANN and PC-ANN for Groups A and B, (A) ANN model for Group A, (B) ANN model for Group B, (C) PC-ANN model for Group A, and (D) PC-ANN model for Group B.

Table 4 – Sensitivity rank of the conventional on-site measurement data in ANN and PC-ANN models from the LH-OAT sensitivity analysis.

	Group A		Group B	
	ANN	PC-ANN	ANN	PC-ANN
Rank 1	Temperature [°C]	pH [–]	Total dissolved solids [g L^{-1}]	Redox potential [mV]
Rank 2	Total dissolved solids [g L^{-1}]	Total dissolved solids [g L^{-1}]	Redox potential [mV]	Electrical conductivity [ms cm^{-1}]
Rank 3	Electrical conductivity [ms cm^{-1}]	Temperature [°C]	Temperature [°C]	pH [–]
Rank 4	Redox potential [mV]	Redox potential [mV]	pH [–]	Temperature [°C]
Rank 5	pH [–]	Electrical conductivity [ms cm^{-1}]	Electrical conductivity [ms cm^{-1}]	Total dissolved solids [g L^{-1}]

models trained with a whole dataset. This result may be attributed to the insignificant relationship between As concentration and the other water quality parameters, and the limited variability of the observed As concentration in Group A. In particular, the prediction for the lower As concentration in Group A is poor, resulting in a lower NSE. In contrast, the model accuracies for Group B in the training and validation steps are satisfactory in terms of NSE (Moriasi et al., 2007).

3.6. *Sensitivity analysis*

The sensitivities of the model outputs (i.e., As concentrations) to water quality parameters were investigated in an attempt to optimize ANN and PC-ANN models using the LH-OAT method. Table 4 presents the sensitivity ranks for the five parameters of ANN and PC-ANN for Groups A and B. For Group A, the most significant parameters of ANN and PC-ANN are temperature and pH, respectively. It seems to be counter-intuitive because pH, TDS, and Eh were identified as significant parameters in previous studies (Buschmann et al., 2007; Amini et al., 2008; Chanpiwat et al., 2011). Conversely, pH is the most sensitive parameter for PC-ANN, followed by TDS and temperature [°C]; this is an acceptable result because neutral to high pH conditions favor As release by promoting desorption processes compared to the predominantly acidic (Buschmann et al., 2007). In addition, pH $\geq$ 7 might possibly enhance the mobilization of As, as explained by Buschmann et al. (2007). The sensitivity results for Group B indicate that Eh is the most sensitive parameter for PC-ANN. This coincides with a previous study on As in Laos (Chanpiwat et al., 2011).

Table 5 – Comparison results between the observed and predicted mean As concentrations (PC-ANN) for each province.

Countries	Province	Observed As [μg L^{-1}]	Predicted As [μg L^{-1}]
Thailand	Suphan Buri	70.0	70.4
Cambodia	Kandal	440.0	325.6
	Phnompen	136.4	127.6
Lao PDR	Vientiane	14.4	24.4
	Borikhamxay	14.0	30.0
	Savannakhet	24.0	6.4
	Saravance	18.8	13.6
	Champasack	40.0	25.6
	Attapeu	11.2	31.6

Also, it is clearly seen that As is negatively correlated with Eh, as shown in Fig. 4. Consequently, even though the predictive power of ANN is sufficient for following variations of the observed As concentration, the role of temperature is theoretically different from that in literature—indeed, it may be caused by a collinearity problem among the input variables. Conversely, the roles of pH and Eh in PC-ANN coincide with preliminary studies, and thereby imply that PC-ANN is the superior model for predicting the As concentrations of groundwater in Cambodia, Laos, and Thailand.

Table 5 compares the averaged observed concentrations and the As concentrations predicted by PC-ANN for each province. In general, the model performances for Cambodia showing a higher As concentration are more accurate than those for the others countries. This may be caused by the training process, which tends to follow a higher observed As, though may not be as useful for lower As concentrations. In other words, the model developed in this study would be more informative for a high-risk area, such as Kandal Province in Cambodia.

4. Conclusions

Groundwater resources are one of the most important components of drinking water supplies, especially in rural areas of Southeast Asian countries including Cambodia, Laos, and Thailand. Over the years, several researchers have attempted to explore the levels of As contamination from various viewpoints. However, few studies have focused on a statistical modeling of As that involved other conventional water quality parameters. As such, the main conclusions drawn from this study are as follows:

1) The poor accuracies of MLR and PCR indicated that linear models for conventional water quality parameters cannot reproduce dynamic variations of observed groundwater As concentrations.
2) The prediction accuracies of ANN and PC-ANN were better than those of MLR and PCR models in both the generation/training and validation steps, showing acceptable accuracies for predicting As concentrations.
3) The results of the sensitivity analysis demonstrated that the predictive power of ANN was satisfactory to follow variations of the observed As, but the roles of the input parameters are theoretically different from those in literature, which might be caused by a collinearity problem among the input variables.

4) Conversely, the roles of pH and Eh in PC-ANN coincided with preliminary studies, and thereby imply that PC-ANN is the superior model for predicting As concentrations of groundwater in Cambodia, Laos, and Thailand.

We expect the PC-ANN model developed in this study to be valuable to not only environmental scientists when designing an efficient As monitoring and removal plan, but also for policy makers and NGOs for establishing effective public health management policies. In particular, because the As testing in a laboratory is a complicated and costly process, the model can be better applied to establish an effective As monitoring and public health management in developing countries. The prediction of As, however, is still a challenging work, especially for developing a reliable model. In this study, the PC-ANN model developed for Cambodia, Laos, and Thailand will be more robust and reliable once new datasets are obtained from other countries. As clearly shown in this study, As contamination could not be explained by a linear combination of conventional water quality parameters. Therefore, this model still needs to be validated by using new datasets in order to more precisely investigate the relationship between As contamination and conventional water quality parameters.

Acknowledgments

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology, Korea (No. 2010-0011822). We would also like to acknowledge the assistance of the Ministry of Land, Transport and Maritime Affairs (MLTMA), and the Korea Meteorological Administration (KMA).

Appendix. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.watres.2011.08.010.

REFERENCES

American Water Works Association (AWWA), 2001. Arsenic Rule, vol. 45. Mainstream.

Amini, M., Mueller, K., Abbaspour, K.C., Rosenberg, T., Afyuni, M., Møller, K.N., Sarr, M., Johnson, C.A., 2008. Statistical modeling of global geogenic fluoride contamination in groundwaters. Environmental Science and Technology 42 (10), 3662–3668.

Bagla, P., Kaise, J., 1996. India's spreading health crisis draws global arsenic experts. Science 274, 174–175.

Berg, M., Tran, H.C., Nguyen, T.C., Pham, H.V., Schertenleib, R., Giger, W., 2001a. Arsenic contamination of groundwater and drinking water in Vietnam: a human health threat. Environmental Science and Technology 35 (13), 2621–2626.

Berg, M., Tran, H.C., Nguyen, T.C., Pham, H.V., Schertenleib, R., Giger, W., 2001b. Magnitude of arsenic pollution in the Mekong and Red River Deltas—Cambodia and Vietnam. Science of the Total Environment 372 (2–3), 413–425.

Berg, M., Luzi, S., Kim, P.T., Viet, P.H., Giger, W., Stuben, D., 2006. Arsenic removal from groundwater by household sand filters: comparative field study, model calculations, and health benefits. Environmental Science and Technology 40 (17), 5567–5573.

Berg, M., Stengel, C., Trang, P.T.K., Hung Viet, P., Sampson, M.L., Leng, M., Samreth, S., Fredericks, D., 2007. Magnitude of arsenic pollution in the Mekong and Red River Deltas—Cambodia and Vietnam. Science of the Total Environment 372 (2–3), 413–425.

Bowerman, B.L., O'Connell, R.T., 1990. Linear Statistical Models: An Applied Approach, second ed. Duxbury Press, Belmont, CA.

Brion, G.M., Lingireddy, S., 1999. A neural network approach to identifying non-point sources of microbial contamination. Water Research 33 (14), 3099–3106.

Buschmann, J., Berg, M., Stengel, C., Sampson, M.L., 2007. Arsenic and manganese contamination of drinking water resources in Cambodia: coincidence of risk areas with low relief topography. Environmental Science and Technology 41 (7), 2146–2152.

Chang, F.J., Kao, L.S., Kuo, Y.M., Liu, C.W., 2010. Artificial neural networks for estimating regional arsenic concentrations in a blackfoot disease area in Taiwan. Journal of Hydrology 388 (1–2), 65–76.

Chanpiwat, P., Sthiannopkao, S., Cho, K.H., Kim, K.W., San, V., Suvanthong, B., Vongthavady, C., 2011. Contamination by arsenic and other trace elements of tube-well water along the Mekong River in Lao RDR. Environmental Pollution 159 (2), 567–576.

Chiew, H., Sampson, M.L., Huch, S., Ken, S., Bostick, B.C., 2009. Effect of groundwater iron and phosphate on the efficacy of arsenic removal by iron-amended biosand filters. Environmental Science and Technology 43 (16), 6295–6300.

Cho, K.H., Kang, J.-H., Ki, S.J., Park, Y., Cha, S.M., Kim, J.H., 2009. Determination of the optimal parameters in regression models for the prediction of chlorophyll-a: a case study of the Yeongsan Reservoir, Korea. Science of the Total Environment 407, 2536–2545.

Gros, P., 1997. Practical linear regression modelling. Basic principles. Validation methods. Oceanis 23 (3), 359–515.

Holmberg, M., Forsius, M., Starr, M., Huttunen, M., 2006. An application of artificial neural networks to carbon, nitrogen and phosphorus concentrations in three boreal streams and impacts of climate change. Ecological Modelling 195 (1–2), 51–60.

Jolliffe, I.T., 2002. Principal Component Analysis, second ed. Springer, New York.

Kocar, B.D., Fendorf, C., 2009. Thermodynamic constraints on reductive reactions influencing the biogeochemical of Arsenic in soils and sediments. Environmental Science and Technology 43 (13), 4871–4877.

Kohnhorst, A., 2005. Arsenic in groundwater in selected countries in South and Southeast Asia: a review. The Journal of Tropical Medicine and Parasitology 28 (2), 73–82.

Kuo, J.T., Hsieh, P.H., Jou, W.S., 2008. Lake eutrophication management modeling using dynamic programming. Journal of Environmental Management 88 (4), 677–687.

Lee, J.H., Yu, M.J., Bang, K.W., Choe, J.S., 2003. Evaluation of the methods for first flush analysis in urban watersheds. Water Science and Technology 48 (10), 167–176.

Luan, F., Liu, H.T., Wen, Y., Zhang, X., 2008. Prediction of quantitative calibration factors of some organic compounds in gas chromatography. Analyst 133 (7), 881–887.

Mac Nally, R., 2002. Multiple regression and inference in ecology and conservation biology: further comments on identifying important predictor variables. Biodiversity and Conservation 11 (8), 1397–1401.

Maier, H.R., Dandy, G.C., 1996. The use of artificial neural networks for the prediction of water quality parameters. Water Resources Research 32 (4), 1013–1022.
Meybeck, M., Carbonnel, J.P., 1975. Chemical transport by Mekong river. Nature 255, 134–136.
Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., Veith, T.L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the ASABE 50 (3), 885–900.
Myers, R., 1990. Classical and Modern Regression with Applications, second ed. Duxbury Press, Boston, MA.
Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I—a discussion of principles. Journal of Hydrology 10 (3), 282–290.
Nguyen, V.L., Ta, T.K.O., Tateishi, M., 2000. Late Holocene depositional environments and coastal evolution of the Mekong River delta, southern Vietnam. Journal of Asian Earth Sciences 18 (4), 427–439.
Polizzotto, M.L., Kocar, B.D., Benner, S.G., Sampson, M., Fendorf, S., 2008. Near-surface wetland sediments as a source of arsenic release to ground water in Asia. Nature 454 (7203), 505–508.
Polya, D.A., Gault, A.G., Bourne, N.J., Lythgoe, P.R., Cooke, D.A., 2003. Coupled HPLC-ICP-MS analysis indicates highly hazardous concentrations of dissolved arsenic species in Cambodian groundwaters. In: Holland, J.G., Tanner, S.D. (Eds.), Plasma Source Mass Spectrometry: Applications and Emerging Technologies. Royal Society of Chemistry, Cambridge, pp. 27–140.
Polya, D.A., Gault, A.G., Diebe, N., Feldman, P., Rosenboom, J.W., Gilligan, E., Fredericks, D., Milton, A.H., Sampson, M., Rowland, H.A.L., Lythgoe, P.R., Middleton, C., Cooke, D.A., 2005. Arsenic hazard in shallow Cambodian groundwaters. Mineralogical Magazine 69 (5), 807–823.
Purkait, B., Kadam, S.S., Das, S.K., 2008. Application of artificial neural network model to study arsenic contamination in groundwater of Malda District, Eastern India. Journal of Environmental Informatics 12 (2), 140–149.
Rawlings, J.O., Pantula, S.G., Dickey, D.A., 1998. Applied Regression Analysis: A Research Tool. Springer, New York.
Riad, S., Mania, J., Bouchaou, L., Najjar, Y., 2004. Rainfall-runoff model using an artificial neural network approach. Mathematical and Computer Modelling 40 (7–8), 839–846.
Rowland, H.A.L., Gault, A.G., Lythgoe, P., Polya, D.A., 2008. Geochemistry of aquifer sediments and arsenic-rich groundwaters from Kandal Province, Cambodia. Applied Geochemistry 23 (11), 3029–3046.
Sarangi, A., Bhattacharya, A.K., 2005. Comparison of artificial neural network and regression models for sediment loss prediction from Banha watershed in India. Agricultural Water Management 78 (3), 195–208.
Smedley, P.L., Kinniburgh, D.G., 2002. A review of the source, behaviour and distribution of arsenic in natural waters. Applied Geochemistry 17 (5), 517–568.
Sousa, S.I.V., Martins, F.G., Alvim-Ferraz, M.C.M., Pereira, M.C., 2007. Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. Environmental Modelling and Software 22 (1), 97–103.
Stanger, G., Truong, T.V., My Ngoc, K.S.L.T., Luyen, T.V., Tuyen, T.T., 2005. Arsenic in groundwaters of the lower Mekong. Environmental Geochemistry and Health 27 (4), 341–357.
Sthiannopkao, S., Kim, K.W., Cho, K.H., Wantala, K., Sotham, S., Sokuntheara, C., Kim, J.H., 2010. Arsenic levels in human hair, Kandal Province, Cambodia: the influences of groundwater arsenic, consumption period, age and gender. Applied Geochemistry 25 (1), 81–90.
Sun, G., Liu, J., Luong, T.V., Sun, D., Wang, L., 2002. Endemic Arsenicosis: A Clinical Diagnostic with Photo Illustrations. UNICEF East Asia and Pacific Regional Office, Bangkok, Thailand, ISBN 974-685-040-7.
Ta, T.K.O., Nguyen, V., Tateishi, M., Kobayashi, I., Saito, Y., 2001. Sedimentary facies, diatom and foraminifer assemblages in a late Pleistocene–Holocene incised-valley sequence from the Mekong River Delta, Bentre Province, southern Vietnam: the BT2 core. Journal of Asian Earth Sciences 20 (1), 83–94.
Tayfur, G., Swiatek, D., Wita, A., Singh, V.P., 2005. Case study: finite element method and artificial neural network models for flow through Jeziorsko Earthfill dam in Poland. Journal of Hydraulic Engineering 131 (6), 431–440.
Tetsuro, A., Takashi, K., Junko, F., Reiji, K., Tu, B.M., Pham, T.K.T., Hisato, I., Annamalai, S., Pham, H.V., Shinsuke, T., 2006. Contamination by arsenic and other trace elements in tube-well water and its risk assessment to humans in Hanoi, Vietnam. Environmental Pollution 139 (1), 95–106.
Wen, C.G., Lee, C.S., 1998. A neural network approach to multiobjective optimization for water quality management in a river basin. Water Resources Research 34 (3), 427–436.