Journal of Environmental Management 151 (2015) 87–96

Contents lists available at ScienceDirect

Journal of Environmental Management

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

Modeling water quality in an urban river using hydrological factors – Data driven approaches

Fi-John Chang [a, *], Yu-Hsuan Tsai [a], Pin-An Chen [a], Alexandra Coynel [b], Georges Vachaud [c]

[a] *Department of Bioenvironmental Systems Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei 10617, Taiwan, ROC*
[b] *Laboratoire d'Environnements et Paléoenvironnements Océaniques et Continentaux, University Bordeaux 1, UMR EPOC, France*
[c] *Laboratoire Transferts en Hydrologie et Environnement, LTHE, UMR 5564 CNRS-IRD-UJF, Grenoble, France*

ARTICLE INFO

Article history:
Received 9 May 2014
Received in revised form
10 November 2014
Accepted 4 December 2014
Available online 24 December 2014

Keywords:
Artificial neural network (ANN)
Water quality
River basin management
Ammonia nitrogen (NH_3–N)
Nonlinear autoregressive with exogenous input (NARX) network
Gamma test

ABSTRACT

Contrasting seasonal variations occur in river flow and water quality as a result of short duration, severe intensity storms and typhoons in Taiwan. Sudden changes in river flow caused by impending extreme events may impose serious degradation on river water quality and fateful impacts on ecosystems. Water quality is measured in a monthly/quarterly scale, and therefore an estimation of water quality in a daily scale would be of good help for timely river pollution management. This study proposes a systematic analysis scheme (SAS) to assess the spatio-temporal interrelation of water quality in an urban river and construct water quality estimation models using two static and one dynamic artificial neural networks (ANNs) coupled with the Gamma test (GT) based on water quality, hydrological and economic data. The Dahan River basin in Taiwan is the study area. Ammonia nitrogen (NH_3–N) is considered as the representative parameter, a correlative indicator in judging the contamination level over the study. Key factors the most closely related to the representative parameter (NH_3–N) are extracted by the Gamma test for modeling NH_3–N concentration, and as a result, four hydrological factors (discharge, days w/o discharge, water temperature and rainfall) are identified as model inputs. The modeling results demonstrate that the nonlinear autoregressive with exogenous input (NARX) network furnished with recurrent connections can accurately estimate NH_3–N concentration with a very high coefficient of efficiency value (0.926) and a low RMSE value (0.386 mg/l). Besides, the NARX network can suitably catch peak values that mainly occur in dry periods (September–April in the study area), which is particularly important to water pollution treatment. The proposed SAS suggests a promising approach to reliably modeling the spatio-temporal NH_3–N concentration based solely on hydrological data, without using water quality sampling data. It is worth noticing that such estimation can be made in a much shorter time interval of interest (span from a monthly scale to a daily scale) because hydrological data are long-term collected in a daily scale. The proposed SAS favorably makes NH_3–N concentration estimation much easier (with only hydrological field sampling) and more efficient (in shorter time intervals), which can substantially help river managers interpret and estimate water quality responses to natural and/or manmade pollution in a more effective and timely way for river pollution management.

1. Introduction

Water quality has deteriorated in most of the major rivers in western Taiwan over the past decades in consequence of urbanization, industrialization and population growth along rivers. Seasonal variations of river flows have also undergone drastic changes due to hydrological and geological characteristics of river basins. Pollutants leak from various sources and accumulate with sediments in river beds. In drought seasons, water levels in general are very low and flows barely occur in river channels. As a consequence, river pollution becomes even worse. On the other hand, during wet seasons, sudden changes in river flow may rinse river beds, deposit particles of sediments and contaminants, and thus river water quality seriously deteriorates and ecosystems encounter huge impacts (Ko et al., 2010). Hydrological characteristics significantly influence the ecological sustainability of aquatic river systems and cause heavy casualties on fishes, shellfishes and mollusks in downstream industries and coastal cultivation in

[*] Corresponding author. Tel.: +886 2 33663452; fax: +886 2 23635854.
E-mail address: changfj@ntu.edu.tw (F.-J. Chang).

http://dx.doi.org/10.1016/j.jenvman.2014.12.014
0301-4797/© 2014 Elsevier Ltd. All rights reserved.

Taiwan (Chang et al., 2013). Water pollution is a crucial problem in the Dahan River (our study area) because many industrial facilities as well as densely populated cities are located along the river in recent decades and uneven dilution as well as transportation of pollutants frequently occurs in the water body of the river. The incessant accumulation of pollutants and nutrients deteriorates river water quality and exhausts dissolved oxygen in the river (Chiu, 2011). Yang et al. (2009) applied the WASP/EUTOR model to evaluating a number of alternatives on wastewater management in the downstream of the Dahan River for river restoration with an improvement on the assimilative capacity of biochemical oxygen demand and dissolved oxygen.

Water quality models are useful tools for estimating the impacts and risks of chemical pollutants in a water body (Chapra, 2008; Feng et al., 2013). Water quality models can be classified into physically-based or statistical approaches. Physically-based approaches have progressively proved successful and useful in learning the mechanisms of underlying processes; nevertheless, they are usually site-specific and require substantially detailed water quality measurements and/or extensive surveys for calibration, which bear certain time and budget limitation. With the development of model theory and the fast-updated computer techniques, more water quality models have been explored with various statistical methods to overcome data scarcity and simultaneously increase model reliability. Either statistical approaches, such as linear regression (Rothwell et al., 2010); factor analysis technique (Ouyang et al., 2000), or artificial neural networks (ANNs) models based on data driven techniques (Unwin et al., 2010; He et al., 2011) can be applied to the monitored time-series of hydrological and water quality measurements for simulation and/or prediction purposes. ANNs are computational techniques inspired by the brain and nerve systems in biological organisms, and they can tackle large-scale complex problems. ANNs have also been applied with success to diverse fields of environment sciences (Coz et al., 2009; McNamara et al., 2008; Yesilnacar et al., 2008; Singh et al., 2009; Chang et al., 2010; Giri et al., 2011; Hattab et al., 2013; Jiang et al., 2013; Tsai. et al., 2014). A majority of studies were dedicated to exploring the applicability of static ANNs, such as the back propagation neural network (BPNN) and the adaptive network-based fuzzy inference system (ANFIS). Nevertheless, the natural characteristics of hydrogeological processes are complex and dynamic. Static neural networks might fail to properly predict the dynamical features of hydrogeological processes, such that the delivered relationship might be simply the possible impacts of factors on the temporal characteristics of local environments (Chang et al., 2013; Chen et al., 2013). Consequently, the comprehensive analysis on the dynamic features of hydrogeological processes and the modeling of tempo-spatial water quality variations remain a great challenge that needs to be overcome.

Environmental sampling is very complicated, laborious, costly and time-consuming. It is unlikely to have continuous long-term water-quality time series data with complete properties at all sampling locations in a river system. Another great challenge for river managers in pollution assessment is the investigation of pollution patterns with high complexity, dynamism and non-linearity in both spatial and temporal scales (Carafa et al., 2007). Management tools require predictive methods or models to relate water quality with hydrological responses, catchment characteristics and/or human activities. We attempt to estimate water quality concentration in shorter time intervals of interest, without conducting water quality field sampling. The proposed scheme comprises artificial neural networks, factor selection and statistics techniques for a comprehensive assessment of river water quality in responses to natural and human activities over the study basin.

2. Methods

This study formulates a systematic analysis scheme (SAS) for assessing the spatio-temporal water quality by artificial intelligence and statistics techniques based on water quality, hydrological and economic data (Fig. 1). A preliminary analysis is conducted through the correlation coefficient analysis to identify the representative water quality parameter (i.e., a correlative indicator of the contamination level over the study area). Three ANNs (BPNN-a classical ANN; ANFIS-a neuro-fuzzy network; and NARX-a dynamic ANN) coupled with the Gamma test (for factor selection) and cross-validation (for data scarcity) are used for modeling the concentration of the representative water quality parameter. The merits of the main methodologies are briefly addressed as follows:

2.1. Gamma test (GT) - factor selection

The use of input selection methods assist in selecting the combination of explanatory variables best suit a model. The proper selection of input variables can improve prediction performance and help understand the processes that resulted in the observed data (Guyon and Elisseeff, 2003). The current study involves water quality, hydrological and economic factors with data sets limited in size, and therefore there is a need to utilize effective input selection methods to characterize the appropriate input—output relationships. The GT, presented by Agalbjorn et al. (1997), is used to estimate the noise level in a data set without assuming any parametric form of equations that govern the system. The only requirement is that the system should be governed by a smooth function because the GT will exploit the hypothesized continuity of this governing function. Performing a single GT is a fast procedure, which can provide the noise estimate for each subset (combination) of input variables. If a subset's associated noise estimate (Γ value) is the closest to zero, it can be considered as the “best combination” of inputs. Recent applications noted that ANNs combined with the GT can obtain accurate estimation based on the identified non-trivial

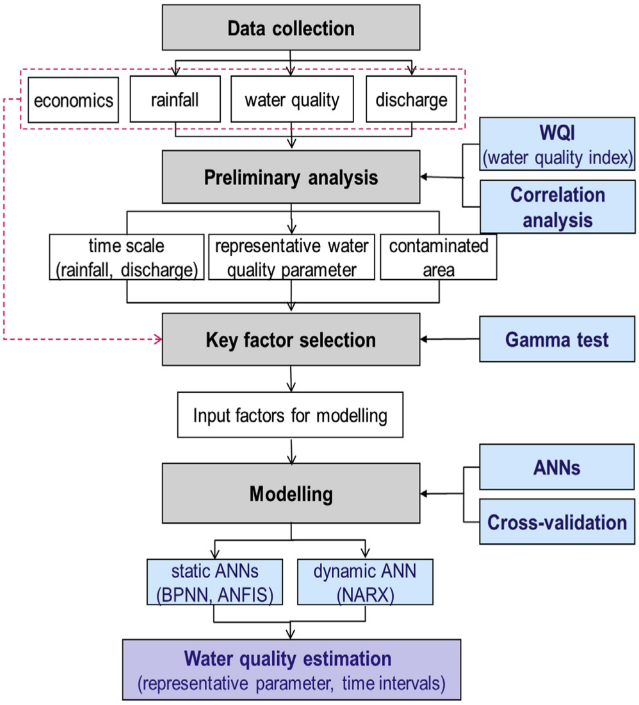

Fig. 1. Study flow of the proposed systematic analysis scheme (SAS).

input variables (Chang et al., 2013; Moghaddamnia et al., 2009; Noori et al., 2011). Therefore, the GT is utilized to extract the key factors for the ANN estimation models in this study.

2.2. Nonlinear autoregressive with eXogenous input (NARX) network – estimation tool

Neural networks possess the ability to approximate nonlinear functions and are valuable tools for modeling time series; however, static neural networks may not establish reliable nonlinear models for predicting dynamical systems and their generalization capability may not satisfy the accuracy and robustness requirement (Shen and Chang, 2012; Chang et al., 2014). Lacking accurate true values makes it difficult to train and construct reliable prediction models. Many engineering problems require models to recover missing data and/or predict into the future without the availability of measurements in the horizon of interest. To assess the models' reliability under this circumstance and determine the most suitable estimation model, two famous static ANNs (BPNN and ANFIS) and one dynamic ANN (NARX) are constructed and validated in this study.

The NARX network proposed by Lin et al. (1996) is an important class of nonlinear discrete-time systems and has two tapped-delay elements produced from input and output layers. Fig. 2 shows the architecture of the NARX network, which consists of three layers (input, hidden and output layers) and the recurrent connections from the output may delay several unit times to form new inputs. This nonlinear system can be mathematically presented by the following equation:

$$z(t) = f[z(t-1), \ldots, z(t-d_z); U(t)] \quad (1)$$

where U(t) and z(t) denote the input vector and output value of the model at a discrete time step t, respectively. And $f(\cdot)$ is the nonlinear function that needs to be approximated by a learning algorithm. When the NARX network needs to be trained, it can be realized in one out of the following two modes. The first mode is the Series-parallel (SP) mode, where the output's regressor in the input layer is formed only by actual values of the system's output, d(t):

$$z(t) = f[d(t-1), \ldots, z(t-d_z); U(t)] \quad (2)$$

The other alternative is the Parallel (P) mode, where estimated outputs are fed back into the output's regressor in the input layer and can be mathematically represented as Eq. (1). It is common that a model adopted to estimate target variables in unrecorded periods often has poor performance because the information of target variables is not always available. Therefore, the NARX network can be trained in the SP mode to construct the relationship between actual and estimated values of the target variable. Then the constructed NARX network in the P mode is applied to the unrecorded period for improving estimation performance with the recurrent information (the estimated values derived from the model). This approach would enhance the estimation accuracy and has practical meaning and functions when dealing with the estimation of target variables in unrecorded periods.

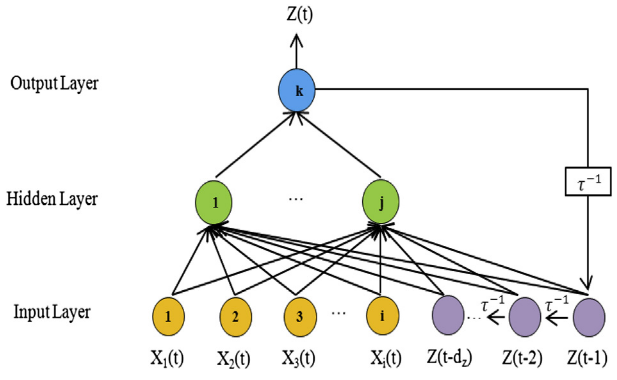

Fig. 2. Architecture of the NARX network without recurrent connections from input-delay terms.

2.3. Cross validation - tackling data scarcity

With data sets limited in size, cross-validation, which partitions observed data into training and testing sets, is commonly used to obtain a reliable estimate of the test error for performance estimation or for use as a model selection criterion. For the k-fold cross-validation, the first step is to assign a model parameter setting (i.e., the initial weights, the epoch number, the number of neurons in the hidden layer and the output-memory orders of the ANN), and then the original sample is partitioned into k subsamples. Among the k subsamples, a single subsample is retained as validation data and the remaining k-1 subsamples are used as training data. The cross-validation process is then repeated k times, with each of the k subsamples being used exactly once as validation data. Cross validation can produce a low-bias estimator for the generalization properties of statistical models, and therefore provides a sensible criterion for model selection and performance comparison, especially for samples that are hazardous, costly or difficult to collect, such as the water quality data collected in this study.

3. Application

The Dahan River has encountered water deterioration, particularly in the mid- and downstream zones in recent years, as a cost of rapid urbanization and industrialization. This study attempts to highlight the representative water quality parameter of this area and reliably estimate its concentration with relatively effortless factors for environmental assessment and pollution management.

3.1. Study area

The Dahan River is located in the upstream of the Danshuei River in northwestern Taiwan (Fig. 3). It is 135 km long and occupies a catchment area of 1163 km^2 with complex land uses involving industrialization, agriculture, urbanization, native vegetation and conservation. Its river basin embraces the densely populated New Taipei City and the Taoyuan County, in which wastewater of households, livestock husbandries and industrial plants infiltrates into the river basin for decades. The Dahan River has become one of the most contaminated rivers in northern Taiwan. Several important hydraulic facilities were built along the river: the Shihmen Reservoir (upstream); the Yuanshan Weir (midstream); and the Bansin Water Intake Plant (downstream). The Yuanshan Weir impounds water from the Shihmen Reservoir and releases water to the Bansin Water Intake Plant for water regulation purpose.

3.2. Data collection and pre-processing

Government-owned water quality monitoring stations have been set up along the main rivers in Taiwan for decades. Various water quality surveys on (heavy) metals and nutrients in rivers and reservoirs have been carried out by the Environmental Protection Administration of Taiwan (TWEPA) since 2002. Water quality data for use in this study were collected at seven water quality

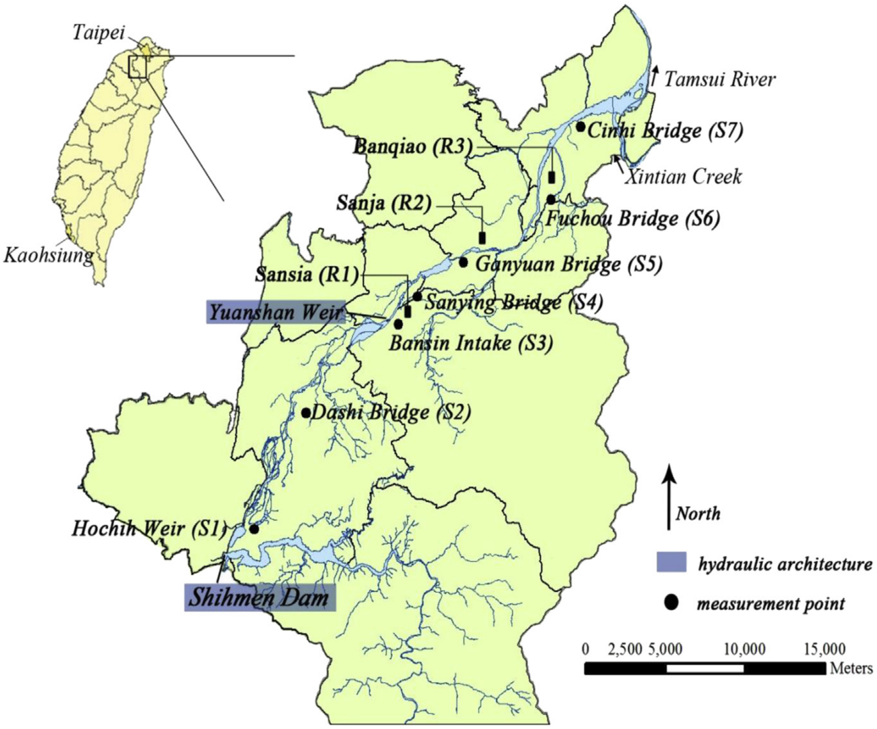

Fig. 3. Dahan River basin with seven water quality monitoring stations and three rainfall gauge stations: upstream zone (S1–S3); and downstream zone (S4–S7).

monitoring stations (S1–S7 in Fig. 3) from June 2002 to June 2012. Monthly data are in association with nine water quality parameters: acidity (pH); electro conductivity (EC); dissolved oxygen (DO); biochemical oxygen demand (BOD); chemical oxygen demand (COD); suspended solid (SS); coliform group (Coliform); ammonia nitrogen (NH_3–N); and water temperature (temp). Quarterly data correspond to five water quality parameters: total organic carbon (TOC); total phosphorous (TP); total nitrogen (TN); nitrate nitrogen (Nitrate); and nitrite nitrogen (Nitrite). Table 1 displays the preliminary statistics of water quality data, which indicate the water quality of the water body varies greatly according to the large standard deviations and the big differences between maximum and minimum concentrations of these parameters.

In addition to water quality parameters, hydrological and economic factors are incorporated into this study for a broad assessment on the influential factors of water quality in the study area. Hydrological factors consist of the discharge from the Yuanshan Weir and the rainfall of three gauge stations (R1-R3) nearby the Dahan River channel (Fig. 3). Discharge data were obtained from the Bansin Water Intake Plant of Taiwan Water Corporation, while rainfall data were collected from the Central Weather Bureau (CWB) and the Water Resources Agency (WRA), Taiwan. Subject to the availability of discharge data collected at the Yuanshan Weir (2004–2011), the modeling period is 2004–2011, instead of 2002–2012 (the period of available water quality data). In the preprocessing stage, rainfall and discharge data are designed into various time series to explore their time lag effects on the representative water quality parameter.

Table 1
Basic statistics of water quality data obtained from seven monitoring stations along the Dahan River (June 2002–June 2012).

Parameters	unit	min	max	Mean	SD[a]	Data quantity
Monthly						
pH		6.4	10.4	7.9	0.5	840
EC	μmho/cm 25 °C	146	1730	411	241	840
DO	mg/l	0	15.8	6.7	3.2	840
BOD	mg/l	1	34.9	4.5	5	840
COD	mg/l	4	551	21.4	36.4	840
SS	mg/l	2.2	24,600	493	1977	840
Coliform	CFU/100 ml	10	68,000,000	558,049	3,003,180	840
NH_3–N	mg/l	0	18.8	1.8	3	840
temp	°C	10	34.7	22.8	5.3	840
Quarterly						
TOC	mg/l	0.71	16.3	4.07	3.09	203
TP	mg/l	0.006	6.88	0.402	0.739	280
TN	mg/l	0.3	21.12	3.92	4.39	126
Nitrate	mg/l	0.01	3.01	0.79	0.58	280
Nitrite	mg/l	0.001	0.942	0.122	0.161	280

[a] Standard deviation.

3.3. Model construction

Fig. 1 illustrates the flowchart of this study. In the preliminary analysis, the representative water quality parameter (i.e., a correlative indicator in judging the contamination level over the study area) is identified by the correlation coefficient analysis on water quality parameters and hydrological factors, respectively. Then, two static ANNs (BPNN and ANFIS) and one dynamic ANN (NARX) coupled with the GT are used to build the concentration estimation models of the representative water quality parameter. The GT is used to select the most suitable combination of factors from water quality, hydrological and economic fields as model inputs. During model construction, each neural network is calibrated by a 14-fold cross validation in consideration of data scarcity (96 data in total). The BPNN and the NARX are trained with the Levenberg–Marquardt optimization algorithm while the ANFIS is trained by the Sugeno fuzzy inference system.

3.4. Performance criteria

The performances of estimation models are evaluated by the commonly used measures of goodness-of-fit: RMSE (Root Mean

Square Error), MAE (Mean Absolute Error), and CE (Coefficient of Efficiency), shown as follows:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{N}\left(Y\left(t\right) - \widehat{Y}\left(t\right)\right)^{2}}{N}} \tag{3}$$

$$MAE = \frac{\sum_{t=1}^{N}\left|Y\left(t\right) - \widehat{Y}\left(t\right)\right|}{N} \tag{4}$$

$$CE = 1 - \frac{\sum_{t=1}^{N}\left(Y\left(t\right) - \widehat{Y}\left(t\right)\right)^{2}}{\sum_{t=1}^{N}\left(Y(t) - \overline{Y}(t)\right)^{2}} \tag{5}$$

where $\widehat{Y}(t)$ is the estimated value, *Y(t)* is the observed value, and *n* is the number of data points. *Y(t-1)* is the estimated value of the preceding time step.

4. Results and discussion

Based on land-use morphology and water contamination level, the study catchment is divided into two zones: the upstream zone (from the Shihmen Reservoir to the Yuanshan Weir; S1–S3); and the downstream zone (from the Yuanshan Weir to the confluence point of the Dahan River and the Xindian River; S4–S7). Fig. 4 shows the water quality of the study area in terms of Water Quality Index (WQI), which is adopted by the TWEPA for river quality assessment in Taiwan. The downstream zone obviously suffers from poor water quality. Therefore, this study will focus in details on the downstream zone for modeling water quality.

4.1. Identification of the representative water quality parameter through correlation analysis

4.1.1. Inter-correlation of water quality parameters

The correlation coefficient analysis is conducted on nine monthly water quality parameters and five quarterly ones, respectively. Table 2 summarizes the correlation results at the downstream zone. For monthly water quality parameters, the parameter pairs of (NH_3–N, DO), (NH_3–N, BOD), (NH_3–N, EC), (BOD, DO) and (EC, DO) occupy top five higher correlation. The correlation result of (NH_3–N, DO) conforms to the biochemical phenomenon: the worse the water quality is, the less the oxygen dissolves in water. NH_3–N not only occupies top three correlation but also significantly correlates with DO, BOD and EC (absolute CC > 0.656), which implies NH_3–N should play a significant correlative role in the deterioration degree of the water body. For quarterly water quality parameters, the parameter pairs of (NH_3–N, TOC), (NH_3–N, TN) and (TOC, BOD) have similar high correlations (CC > 0.8). The important nitrogen pollutants, NH_3–N and TN, produce high positive correlations with EC, BOD and TP. NH_3–N-related pairs also produce the highest correlations, which also imply NH_3–N should play a

Table 2
Correlation of water quality parameter pairs in the downstream zone of the Dahan River (June 2002–June 2012).

Time scale	Parameter pair	Correlation Coefficient	Ranking
Monthly	NH_3–N, DO	−0.762	1
	NH_3–N, BOD	0.688	2
	NH_3–N, EC	0.656	3
	BOD, DO	−0.653	4
	EC, DO	−0.653	4
	EC, BOD	0.571	6
	BOD, COD	0.551	7
	SS, COD	0.483	8
	SS, pH	0.415	9
	EC, COD	0.395	10
	COD, DO	−0.387	11
	NH_3–N, Coliform	0.352	12
	NH_3–N, COD	0.340	13
Quarterly	TOC, NH_3–N	0.846	1
	TN, NH_3–N	0.832	2
	TOC, BOD	0.819	3
	TN, DO	−0.772	4
	TOC, TN	0.732	5
	TN, EC	0.724	6
	TP, SS	0.681	7
	TOC, DO	−0.681	7
	TN, BOD	0.671	9
	TOC, EC	0.658	10
	TP, TN	0.638	11
	TP, COD	0.630	12
	TP, DO	−0.446	13

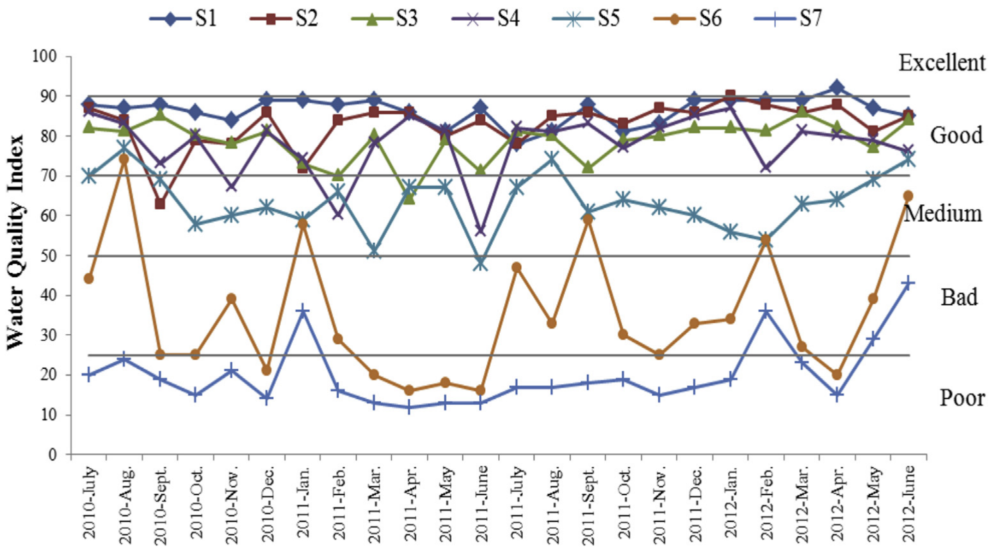

Fig. 4. Monthly WQI indexes (without TP) calculated from seven water quality monitoring stations (S1–S7) along the Dahan River (July 2010–June 2012).

significant correlative role in the deterioration degree of the water body.

4.1.2. Determination of representative water quality parameter

NH_3–N is known to exist in fertilizers, septic system effluent and animal wastes. High levels of NH_3–N in its un-ionized form can be toxic to aquatic organisms, which raises a concern. Conversion of NH_3–N to nitrite nitrogen by nitrification consumes large amounts of oxygen, and thus aquatic organisms can be destroyed due to the lowered dissolved oxygen concentrations in water. Taking the correlation results shown in Table 2 into consideration, NH_3–N is considered as the representative water quality parameter that the most significantly correlates with other water quality parameters affecting river water quality in the study area. Fig. 5 exhibits the spatial variances of three nitrogen compounds (NH_3–N, Nitrite and Nitrate) in the study area during 2007 and 2011. A higher percentage of NH_3–N (in red) in a pie-chart indicates an instant leakage of unnitrified pollutants, and the percentage of NH_3–N drastically increases from S5 to S7. It appears that the pollution degrees of these three nitrogen compounds well conform to the degree of economic development expressed by the orange dots (denote industrial facilities and farms).

Previous studies indicated that the removal of the vegetation cover through severe poaching led to an increase in the delivery rate of NH_3–N in surface runoff (Heathwaite et al., 1990), and the ranking of nitrogen contamination levels is related to the regional population and economic development (Jingsheng et al., 2000). Fig. 6 further displays two selected highly NH_3–N-contaminated conditions at the downstream zone during 2004 and 2011. River water quality may dramatically deteriorate during dry periods due to the drop in natural flow and water temperature. The most concerned period in the study area is between September and April (dry period), during which NH_3–N concentrations sometimes reach as high as 16.0 mg/l, far beyond the drinking water quality standard (0.5 mg/l).

4.1.3. Correlation between hydrological factors and water quality parameters

For depicting the time lag phenomena between hydrological factors and water quality parameters, rainfall data collected at rainfall gauge stations (R1-R3) are converted into two time series: previous day (average rainfall of the previous day over R1-R3); and previous 5-day average (average rainfall of the current day and previous four days) prior to the water quality sampling day. Similar arrangement is made for the discharge data of the Yuanshan Weir: current day (the water quality sampling day); previous day; and previous 5-day average (average of the current day and previous four days). In addition, a new factor "days w/o discharge" is introduced to present the time interval between the previous discharge day and the water quality sampling day. As expected, the correlation coefficients (CC) are not high (less than 0.6); consequently parameters with relatively high correlation results are selected. EC, DO and NH_3–N show moderate relations (absolute CCs fall between 0.21 and 0.51) for both rainfall time series. EC, DO, BOD and NH_3–N have higher relationship (absolute CCs fall between 0.4 and 0.6) for two discharge time series (previous day, and previous 5-day average). It can be inferred that the longer the period w/o discharge is, the higher the NH_3–N concentration is in the water body. In brief, rainfall (previous day, previous 5-day average) and discharge (previous day, 5-day average, das w/o discharge) significantly correlate with the fluctuation of NH_3–N concentration, and therefore both hydrological factors can be considered as inputs for modeling.

4.2. Key factor selection through the GT

For building ANN models to estimate NH_3–N concentration, key input factors are selected by the GT in this study. The study area experiences high degrees of urbanization and industrialization, it would be interesting to understand the impacts of economic factors (human activities) on NH_3–N concentration. Therefore, major economic indexes are the first time included at the factor selection

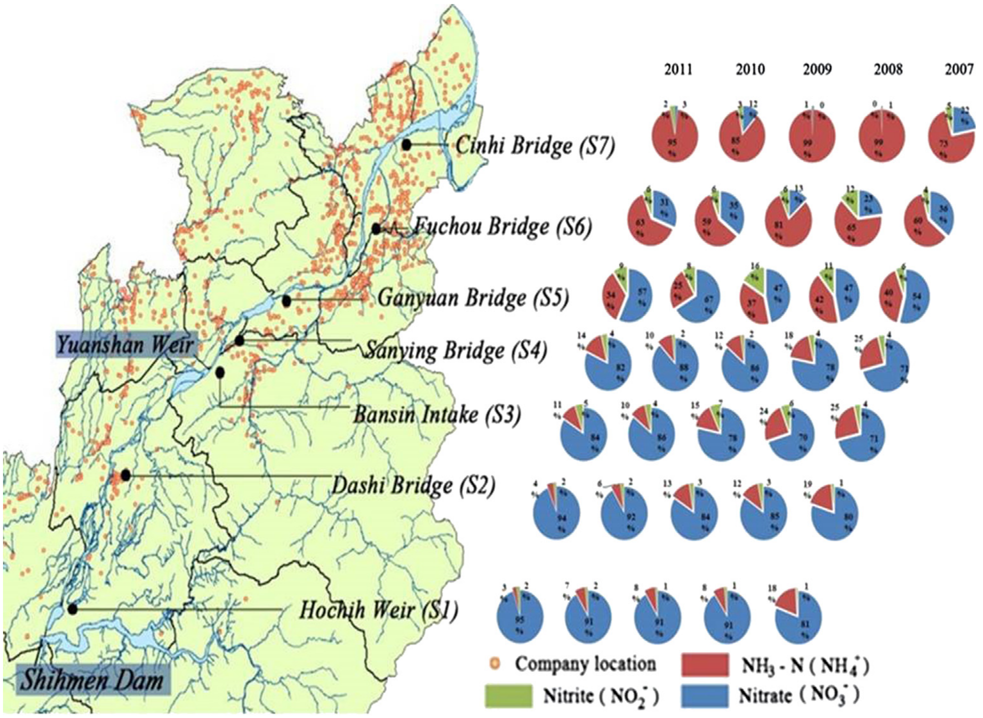

Fig. 5. Spatial variances of three nitrogen compounds (NH_3-N, Nitrite and Nitrate) at seven water quality monitoring stations (2007–2011).

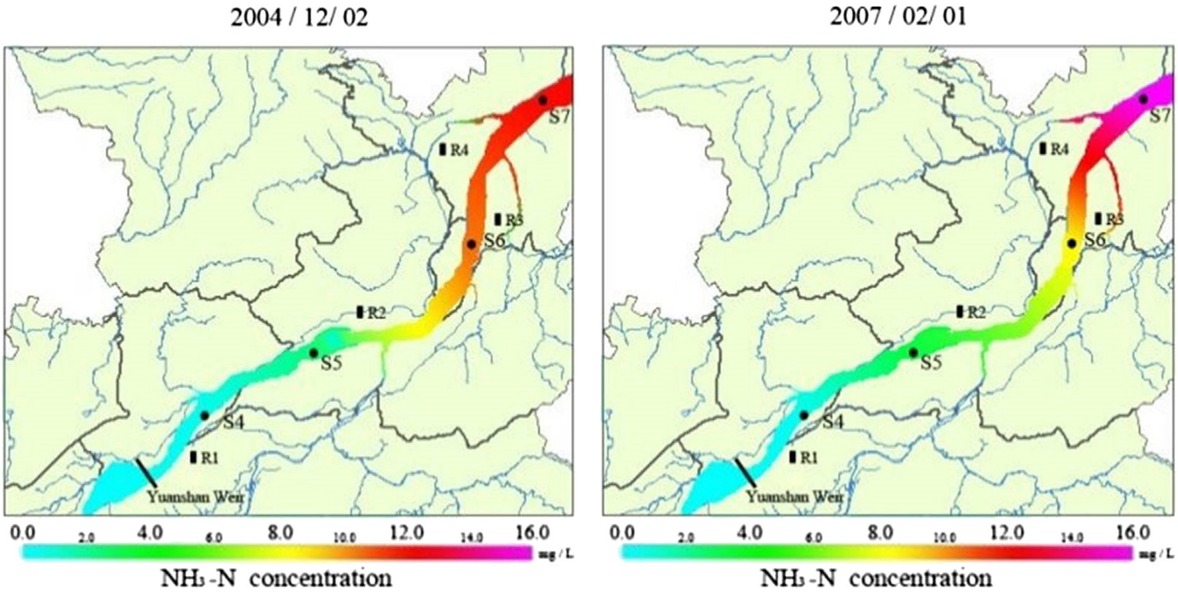

Fig. 6. Selected highly NH_3–N-contaminated conditions at the downstream zone.

stage. The correlation analyses are conducted on thirty-one factors from water quality (9), hydrological (10) and economic (12) fields to reduce the dimension of the GT (2^{31}-1 = 2,147,483,647 input combinations, which is too large). Eleven factors (in italic) bearing the highest correlation coefficients in each fields are filtered out and are investigated further by the GT for obtaining an input combination best suit for modeling (Table 3). It is noted that three economic factors are selected even though their CC values are comparatively low for human impact assessment.

A total of 2047 (2^{11}-1) Γ values corresponding to all possible combinations of factors are calculated. Γ values are sorted in an ascending order, values smaller than the 1st percentile (Γ_1) of all Γ values are defined as the best results of the GT, whereas Γ values bigger than the 99th percentile (Γ_{99}) of all Γ values are defined as the worst results of the GT. By examining the ratio of the occurrence frequency of each factor in the set of best results ($F_\Gamma \leq \Gamma_1$) to that of worse results ($F_\Gamma \geq \Gamma_{99}$), key factors are determined as the factors that produce larger ratios. Fig. 7 shows the GT results, where the blue bars represent the occurrence frequency of a factor in the best results ($F_\Gamma \leq \Gamma_1$) while the red bars represent the occurrence frequency of a factor in the worst results ($F_\Gamma \geq \Gamma_{99}$). Four factors (i.e., discharge of previous day, days w/o discharge, water temperature of previous month, and rainfall of previous day) with the highest ratios are identified as the most significant factors affecting the NH_3–N concentrations. We notice that all the four key factors belong to hydrological factors, which are relatively easier to observe as compared with water quality and/or economic factors. The responses of these four factors to NH_3–N concentrations conform to physical processes (discharge and rainfall makes significant effects on river flow, and thus cause the fluctuation of NH_3–N concentration; water temperature strongly affects ammonia oxidation activity (Groeneweg et al., 1994)), which supports the suitability of the four factors as model inputs.

4.3. *Estimation of NH_3–N concentration by ANNs*

With the previous four selected factors as input variables and the NH_3–N concentrations as the output, this study adopts three ANNs to estimate the average NH_3–N concentration at S4–S7 in the study area. Each neural network is calibrated by a 14-fold cross validation, and the 96 data sets (2004–2011) are allocated: 84 data sets (2004–2010) for model calibration; and the remaining 12 data sets (2011) for model testing. The model performance in the testing stages is listed in Table 4. The results indicate that the NARX network performs the best, which produces the smallest RMSE and MAE values and the highest CE value. To be more specific, the NARX network produces worthy and accurate results when comparing its RMSE (0.386 mg/l) and MAE (0.277 mg/l) with the mean and standard deviation (3.072 ± 1.497 mg l^{-1}) of the averaged NH_3–N concentrations over the downstream zone. We would like to note that the estimation models are constructed based only on four hydrological variables that are long-term collected in a daily, even hourly or minutely, scale, and therefore NH_3–N concentration can be estimated in a much shorter time interval of interest (span from a monthly scale to a daily, even hourly or minutely, scale) through the constructed model for timely water pollution management.

Ultimately, the performance of the individual modeling approaches depends on how well these ANNs can describe the underlying physical process as well as specific data sets, and results are influenced by the affecting factors and conditions associated with the gauge stations in the study area. Fig. 8 shows the estimation time series of NH_3–N concentrations associated with three constructed ANNs. The results demonstrate that (1) all the models, in general, well fit the variability of the NH_3–N concentration; (2) the NARX network more accurately reflects the fluctuations of the NH_3–N concentrations owing to its dynamic feature-recurrent connection; and (3) the NARX network can well catch peak values that mostly occur in dry periods (September–April), which is particularly important to water pollution treatment. In sum, the NARX network coupled with the GT appears most adequate in delineating the water quality with hydrological variables.

5. Conclusion

For efficiently assisting in river pollution management and providing a cost-manpower effective approach to making timely response to natural and/or manmade pollution, this study proposes a systematic analysis scheme (SAS) to first identify the representative water quality parameter by exploring the extensive connections of parameter pairs of water quality data and next develop

Table 3
Correlation results of NH_3–N and factors from water quality, hydrological and economic fields (2004–2011*).

Factor	CC[a]	Factor	CC
Water quality parameters		***Economic indexes***	
(S4–S7; Previous month)		*(Previous month)*	
pH	−0.163	TWSE TAIEX Index [d]	
EC	0.097	Change extent	0.154
DO	−0.006	Monthly trade volume	−0.068
BOD	−0.039	TAIEX-mechatronics engineering	
COD	−0.06	*Change extent (T(ME)-CE)*	0.219
SS	0.102	Monthly trade volume	−0.042
Coliform	0.195	GTSM Index [e]	
NH_3–N	0.222	*Change extent (GTSM CE)*	0.184
Water temperature	**−0.388**	Monthly trade volume	0.035
Hydrological factors		Industrial Production Index	
Rainfall (R1-R4)		Volume	−0.131
Previous day	**−0.266**	Manufacturing industry	−0.127
Previous 5-day average	−0.464	Commodity industry	−0.136
Discharge (Yuanshan Weir)		Chemistry industry	−0.167
Current day	0.039	*Metal & Machinery Industry (IPMMI)*	−0.204
Previous day	**−0.331**	Index of Producers Shipment	
Previous 5-day average	−0.332	Volume	−0.114
Days w/o discharge	**0.593**		
Previous discharge 1[b]	−0.189		
Discharge maintain rate 1[c]	−0.191		
Previous discharge 2[b]	−0.362		
Discharge maintain rate 2[c]	−0.406		

* Discharge data of the Yuanshan Weir are available in the period of 2004 and 2011. Therefore, the modeling period is limited to 2004–2011, instead of 2002–2012 (the period of available water quality data in this study). Candidate factors for the GT are marked in italic while the selection results of the GT are marked in bold.

[a] Correlation coefficient.

[b] Previous discharge: previous discharge quantity of the Yuanshan Weir on the sampling day of NH_3–N (1: w/current day volume; 2: w/o current day volume).

[c] Discharge maintain rate: previous discharge divided by Days w/o discharge.

[d] TWSE TAIEX Index: Taiwan Stock Exchange Capitalization Weighted Stock Index, compiled by Taiwan Stock Exchange Corporation.

[e] GTSM Index: Gre Tai Securities Market Capitalization Weighted Stock Index.

Table 4
Model performance of NH_3–N concentration at S4–S7 in the testing stages.

Model	Node number	RMSE	MAE	CE
BPNN	4	0.620	0.504	0.829
ANFIS	3	0.753	0.644	0.747
NARX	4	0.386	0.277	0.926

estimation models for the representative parameter by using three ANNs (one dynamic-NARX and two static-BPNN and ANFIS) with one advanced factor selection method based on a number of factors selected from water quality, hydrological and economic domains. The Dahan River is the study area. Significant findings are addressed as follows:

1) Serious contamination problems are found to exist at water quality monitoring stations S5–S7 (the downstream zone) according to WQI results;
2) NH_3–N is identified as the representative water quality parameter most significantly correlative with other water quality parameters responsible for water quality deterioration the study area;
3) Four key factors (discharge of previous day, days w/o discharge, water temperature of previous month, and rainfall of previous day) are identified as model inputs by the GT through assessing a large number (2047 in this case) of all possible input combinations in association with nine factors selected from water quality, hydrology and economic fields. It is worth noting that these four inputs consist only of hydrological factors, which shows a more influential role the hydrological factors play;
4) The dynamic NARX network produces the most accurate estimation results for NH_3–N concentration than the two static neural networks. This is mainly because that the NARX is furnished by its recurrent connections, which is more suitable to track the dynamic features of the estimated time series. It also can well captures peak values that mostly occur in dry periods, which is particularly important to water pollution treatment; and
5) The proposed SAS is considered as a breakthrough methodology for reliably modeling NH_3–N concentration, which pivots solely on hydrological data, without the use of water quality sampling

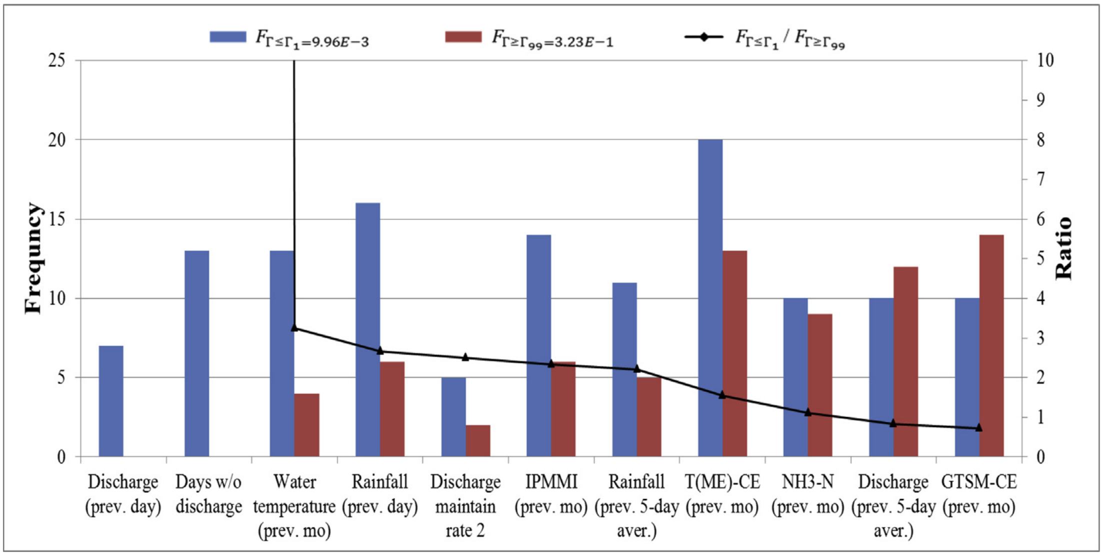

Fig. 7. Determination of key factors by the GT.

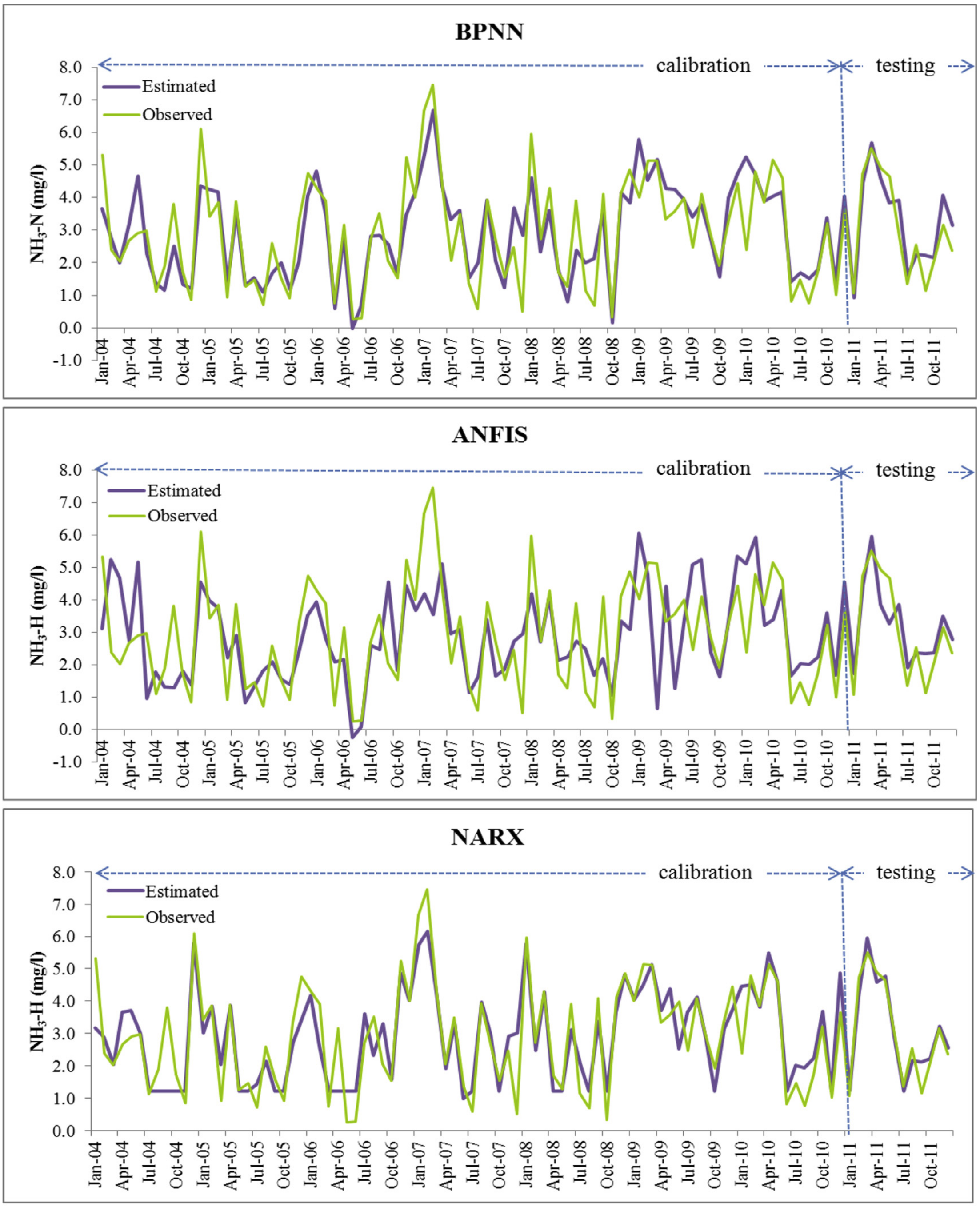

Fig. 8. Performance comparison of the regional NH_3–N concentrations.

data. The scheme also implies that estimation can be made in a much shorter time interval of interest (span from a monthly scale to a daily scale) because hydrological data are usually measured in a daily scale.

The proposed analytical scheme favorably estimates NH_3–N concentration in a much easier (with only hydrological field sampling) and more efficient (in shorter time intervals) way and can be appropriately and practically applied to other water quality parameters and/or study areas of interest, which can substantially help river managers timely interpret and estimate water quality responses to natural and/or manmade pollution for river pollution management.

Acknowledgements

This study was partially funded by the Ministry of Science and Technology, Taiwan, ROC. (Grant numbers: 101-2923-B-002-001-MY3).

References

Agalbjorn, S., Koncar, N., Jones, A.J., 1997. A note on the gamma test. Neural Comput. Appl. 5, 131–133.

Carafa, A., Lumini, E., Bianciotto, V., Bonfante, P., Duckett, J.G., 2007. Glomeromycotean associations in liverworts: a molecular cellular and taxonomic analysis. Am. J. Bot. 94 (11), 1756–1777.

Chang, F.J., Chen, P.A., Lu, Y.R., Huang, Eric, Chang, K.Y., 2014. Real-time multi-step-ahead water level forecasting by recurrent neural networks for urban flood control. J. Hydrol. 517, 836–846.

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. J. Hydrol. 388, 65–76.

Chang, F.J., Tsai, W.P., Chen, H.K., Yam, R.S.W., Herricks, E.E., 2013. A self-organizing radial basis network for estimating riverine fish diversity. J. Hydrol. 476, 280–289.

Chapra, S.C., 2008. Surface Water-quality Modeling. Waveland Press.

Chen, P.A., Chang, L.C., Chang, F.J., 2013. Reinforced recurrent neural networks for multi-step-ahead flood forecasts. J. Hydrol. 497, 71–79.

Chiu, H.H., 2011. The Study of Water Quality of the Da-han Stream and Danshuei Estuary. A Thesis Submitted to Department of Marine Environmental Informatics College of Engineering Science and Resource. National Taiwan Ocean University.

Coz, E., Gómez-Moreno, F.J., Pujadas, M., Casuccio, G.S., Lersch, T.L., et al., 2009. Individual particle characteristics of North African dust under different long-range transport scenarios. Atmos. Environ. 43, 1850–1863.

Feng, Y., Barr, W., Harper Jr., W.F., 2013. Neural network processing of microbial fuel cell signals for the identification of chemicals present in water. J. Environ. Manag. 120 (15), 84–92.

Giri, A.K., Patel, R.K., Mahapatra, S.S., 2011. Artificial neural network (ANN) approach for modelling of arsenic (III) biosorption from aqueous solution by living cells of Bacillus cereus biomass. Chem. Eng. J. 178, 15–25.

Groeneweg, J., Sellner, B., Tappe, W., 1994. Ammonia oxidation in nitrosomonas at NH_3 conentrations near Km: effects of pH and temperature. Water Res. 28 (12), 2561–2566.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.

Hattab, N., Hambli, R., Motelica-Heino, M., Mench, M., 2013. Neural network and Monte Carlo simulation approach to investigate variability of copper concentration in phytoremediated contaminated soils. J. Environ. Manag. 129 (15), 134–142.

He, B., Oki, T., Sun, F., Komori, D., Kanae, K., Wang, T., Kim, H., Yamazaki, D., 2011. Estimating monthly total nitrogen concentration in streams by using artificial neural network. J. Environ. Manag. 92 (1), 172–177.

Heathwaite, A.L., Burt, T.P., Trudgill, S.T., Thornes, J.B., 1990. The effect of land use on nitrogen, phosphorus and suspended sediment delivery to streams in a small catchment in southwest England. In: Vegetation and Erosion. Processes and Environments. Wiley, Chichester, pp. 161–177.

Jiang, Y., Nan, Z., Yang, S., 2013. Risk assessment of water quality using Monte Carlo simulation and artificial neural network method. J. Environ. Manag. 122 (15), 130–136.

Jingsheng, C., Xuemin, G., Dawei, H., Xinghui, X., 2000. Nitrogen contamination in the Yangtze River system, China. J. Hazard. Mater. 73 (2), 107–113.

Ko, C.H., Chang, F.C., Lee, T.M., Chen, P.Y., Chen, H.H., Hsieh, H.L., Guan, C.Y., 2010. Impact of flood damage on pollutant removal efficiencies of a subtropical urban constructed wetland. Sci. Total Environ. 408 (20), 4328–4333.

Lin, T., Horne, B.G., Tino, P., Giles, C.L., 1996. Learning long-term dependencies in NARX recurrent neural networks. IEEE Trans. Neural Netw. Learn. Syst. 7 (6), 1424–1438.

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. Hydrol. Process 22, 3294–3309.

Moghaddamnia, A., Ghafari, M., Piri, J., Amin, S., Han, D., 2009. Evaporation estimation using artificial neural networks and adaptive neuro-fuzzy inference system techniques. Adv. Water Resour. 32, 88–97.

Noori, R., Karbassi, A., Moghaddamnia, A., Han, D., Zokaei-Ashtiani, M.H., Farokhnia, A., Gousheh, M.G., 2011. Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction. J. Hydrol. 401, 177–189.

Ouyang, Y., Nkedi-Kizza, P., Wu, Q.T., Shinde, D., Huang, C.H., 2000. Assessment of seasonal variations in surface water quality. Water Res. 40 (20), 3800–3810.

Rothwell, P.M., Wilson, M., Elwin, C.E., Norrving, B., Algra, A., Warlow, C.P., Meade, T.W., 2010. Long-term effect of aspirin on colorectal cancer incidence and mortality: 20-year follow-up of five randomised trials. Lancet 376, 1741–1750.

Shen, H.Y., Chang, L.C., 2012. On-line multistep-ahead inundation depth forecasts by recurrent NARX networks. Hydrol. Earth Syst.Sci 9 (10). Discussions.

Singh, K.P., Basant, A., Malik, A., Jain, G., 2009. Artificial neural network modeling of the river water quality—a case study. Ecol. Model 220 (6), 888–895.

Tsai, M.J., Abrahart, R.J., Mount, N.J., Chang, F.J., 2014. Including spatial distribution in a data-driven rainfall-runoff model to improve reservoir inflow forecasting in Taiwan. Hydrol. Process 28, 1055–1070.

Unwin, M., Snelder, T., Booker, D., Ballantine, D., Lessard, J., 2010. Predicting water quality in New Zealand rivers from catchment-scale physical, hydrological and land use descriptors using random forest models. Tech. Rep. Ministry Environ. CHC2010-037 (New Zealand).

Yang, C.P., Lung, W.S., Liu, J.H., Hsiao, W.P., 2009. Establishment and application of water quality model of Hypoxic stream. J. Taiwan Agric. Eng. 55 (1), 27–39.

Yesilnacar, M.I., Sahinkaya, E., Naz, M., Ozkaya, B., 2008. Neural network prediction of nitrate in groundwater of Harran Plain Turkey. Environ. Geol. 56, 19–25.