JOURNAL OF
ENVIRONMENTAL HYDROLOGY

The Electronic Journal of the International Association for Environmental Hydrology
On the World Wide Web at http://www.hydroweb.com

VOLUME 11 2003

INVERSE MODELING TO IDENTIFY NONPOINT SOURCE POLLUTION USING A NEURAL NETWORK, TAIHU LAKE WATERSHED, CHINA

Iqbal Zaheer
Guangbai Cui

College of Water Resources and Environment
Hohai University, Nanjing, China

Various studies have been carried out for the evaluation of non-point source pollution using physically-based distributed hydrological and water quality models. A number of modeling interactions have been developed using remote sensing, geographical information systems, best management practices, decision support systems, and water quality modeling tools, for the identification and quantification of non-point sources of pollution in the watersheds of lakes and rivers. In the most recent artificial neural network applications, an intelligence system is commonly used for the management of the dynamic and complex nature of watersheds. In this paper, the author has proposed an inverse modeling approach, using artificial intelligence for the identification of non-point source pollution based on pollution indicators in storm water and agriculture runoff. The study is carried out in the Xishan county sub basin of the Taihu Lake watershed. A back propagation neural network model has assisted the development of an inverse modeling system to identify pollution sources beyond the presence of pollution indicators.

INTRODUCTION

In the past, water resources engineers used various methods of lumped and distributed mathematical modeling, geographical information systems (GIS) and neural nets to tackle the problems of non-point source pollution (NPSP) as part of the watershed management process. The development of population, industry and agriculture has caused an increase of the nutrients nitrogen and phosphorous in the water bodies of Europe and America (Addiscott et al., 1992). Various studies have been carried out for the identification, quantification and modification of NPSPat the watershed scale (Cooper and Thomson, 1988; Hopkinson and Vallino, 1995). Regarding quantification of NPSP based on information on land use, demographics, and the hydrological environment, large numbers of urban and agriculture NPSP models have been developed that can be applied to study diffused pollution in the watershed (Giorgini and Zingales et al. 1986, Donigian and Huber 1990, Young et al. 1986, Arnold et al. 1993). Most of these models are composed of digital descriptions of geographical features of land and soil, and are integrated with physically based mathematical techniques describing the process governing return flows, water quality, hydrological features and transport of pollutants in the watershed. Furthermore, modeling interactions of GIS with physical approaches are described in numerous papers (Zaheer and Cui, 2002a).

The Artificial Neural Network (ANN) has been widely used in various studies of surface water pollution control for predicting stream nitrogen concentration (Lek et al., 1999), forecasting raw water quality parameters (Zhang and Stanley, 1997), prediction of water quality parameters (Zaheer and Cui, 2002b), water quality management (Wen and Lee, 1998), and identification of non-point sources of microbial contamination (Brion and Lingireddy, 1999). The ANN may be applied to different kinds of approaches, e.g. pattern classification, interpretation, generalization, or calibration. In this paper, the ANN is applied for pattern classification for the development of an inverse modeling approach. Since Neural Net (NNs) models are applied to identify pollution indicators from storm water of varying watershed features of NPSP, the novel application of ANN can be used for the identification of NPSP using the pollution potential indicator in storm and agriculture runoff. This leads to the application of an inverse modeling approach for the development of watershed management and planning tools.

Neural network capability for watershed management

Surface water quality control is the predominant factor in the watershed management process. The potential of pollutant sources, hydrological parameters and human activities, and their impact on water quality creates complex and dynamic watersheds. The precision in the system is decreased with increase of system complexity until a threshold condition where precision becomes a mutually exclusive characteristic as described in Kosko, (1993). In such a complex system, the ANN is used as an approximation tool rather then a complex mathematical calculation, which results in a ten percent deviation of predicted value from observed data (Lingireddy and Ormsbee, 1998). Using the inverse modeling approach, a neural network could be established for pattern classification to identify NPSP from the pollutant indicators of storm water runoff.

The study is carried out in the Taihu Lake watershed located in Xishan county, China. The problem is focused on storm water runoff from 13 sub-counties of Xishan county having extensive land uses of agriculture (pasturelands and paddy fields), domestic (residential and commercial), and livestock (animals and fisheries). The major pollutant indicator variables are COD and nutrients, including N and P, which result in severe impacts to surface water quality within the watershed.

Using variable pollutant indicator data of a single event in storm runoff, a neural network is trained to find the unique non-point sources of pollution of watershed features.

NEURAL NETWORK SOURCE IDENTIFICATION MODEL

The variable concentration of pollutant indicators in surface runoff are released from watershed land uses of agriculture, domestic use, and livestock. This study is designed to train a neural network to recognize the pollutant indicators in surface runoff for source identification.

Topography of the area

The Xishan county part of the Taihu Lake watershed covers a total area of 878 km^2, consisting of about 409 km^2 made up of 13 counties with 31% agriculture, 17% domestic use and 6% livestock. Most of the area is a plain having extensive agriculture in paddy fields, pasture lands, grasslands and forests, where most of the pollutants are concentrated. The rapid economic growth due to agriculture, domestic, industrial and livestock development is causing severe impairment of surface runoff. Three major tributaries of Taihu Lake in Xishan county are the Xi Bei Yuan , Jiu Li, and Bodugang Rivers.(Chen and Li, 1998). The topographic map of Taihu watershed is shown in Figure 1.

Development of input data

Data used in this study is simulated using a one dimensional water quality model for the period January 2001 to December 2001 (Zaheer and Cui, 2002c). Data consist of four sets of pollutant indicator concentrations of COD, N and P, collected from various land use features of the 13 sub-counties. Two sets are used for training, and two sets are used for model prediction. Moreover, the nutrient concentration of total N, total P and COD were measured monthly for one year from storm runoff corresponding to various land use features of all counties, including domestic, terrestrial (animal husbandry) and aquatic (fisheries) livestock. For the NNs model simulation, mean annual concentration is computed based on the arithmetic mean of all the values for a given sampling site for a year of sampling. In the present study, we consider pollutant indicators as the independent variable including total N (TN), total P (TP) and COD collected randomly from the storm water runoff of various land use features. Other data categories are processed, which are defined as average annual precipitation (AAP), population density (POD), animal unit density (AUD), and average

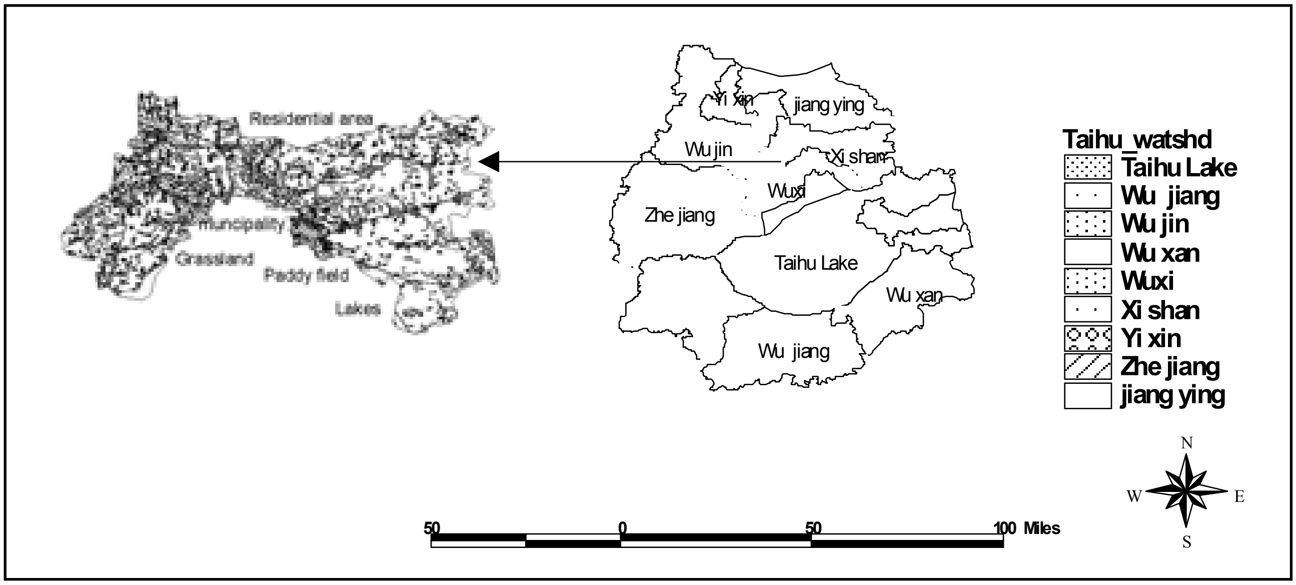

Figure 1. Topographic map of Xishan county in Taihu Lake watershed.

annual surface runoff (ASR) (Table 1). The percentages of land use watershed features considered as dependent variables include agriculture (AGR), domestic (DOM), terrestrial livestock (LS-1) and aquatic livestock (LS-2) (Table 2). In this inverse modeling system, both data sets are used for the training of NNs and model prediction.

Table 1. Input Data for the Training and Prediction of Neural Nets

Name of Sub-counties	Pollutants Indicators from Strom runoff of various land use features selected randomly			AAP	AGD Acres	POD	AUD		ASR
							1000 kg		
	TN	TP	COD	(mm)	Acres	10,000	LS-1	LS-2	(mm)
Cha iao	54.77	0.39	-	943	16905	1.79	69	171	2.24
Hou Qiao	7.38	3.09	36.88	943	17460	1.67	58	1228	2.06
An Zhang	13.03	0.21	25.56	943	34080	2.53	78	392	3.37
Hou Zai	15.67	0.49	-	943	28170	2.70	65	301	2.91
Hong Sheng	108.92	0.77	-	943	21465	1.98	58	223	2.25
Zhang Ling	10.56	4.42	52.78	943	9300	2.39	70	425	3.34
Dong Tin	9.37	0.18	13.56	943	36630	3.05	63	452	2.59
Donghu Tang	14.48	0.31	-	943	29430	2.71	77	700	3.60
Ba Shi	18.64	0.78	-	943	17985	2.18	69	358	2.88
Dongbei tang	62.16	1.71	37.76	943	19185	171	58	795	1.96
Dong Kou	8.35	3.49	9.32	943	18930	1.89	59	1881	2.38
Gan Lu	9.37	0.24	-	943	37290	3.71	68	402	2.34
Gang Xia	120.82	0.86	18.37	943	33525	2.45	57	345	3.89

Table 2. Identified Output and Model Identification

Name of Sub-counties	Land use Features	Identified output	Sources Identification	Name of Sub-counties	Land use Features	Identified output	Sources Identification
Dong Hei tang	AGR	0.00	0.08	Cha Qiao	AGR	0.00	0.09
	DOM	0.25	0.16		DOM	0.25	0.21
Ba Shi	AGR	0.00	0.00		LS-2	0.75	0.70
	DOM	0.25	0.25	An Zheng	AGR	0.00	0.00
Zhang Jing	AGR	0.00	0.08		LS-2	0.75	0.75
	DOM	0.25	0.23	Hou Qiao	AGR	0.00	0.05
Dung Hu yang	AGR	0.00	0.02		DOM	0.25	0.00
	DOM	0.25	0.01	Hong Shang	AGR	0.00	0.09
	LS-2	0.75	0.56		LS-1	0.50	0.45
Gang Xia	AGR	0.00	0.00	Dang Kou	AGR	0.00	0.12
	LS-1	0.50	0.50		DOM	0.50	0.20
Gan Lu	AGR	0.00	0.08		DOM	0.25	0.25
	DOM	0.25	0.25	Hou Zhai	AGR	0.00	0.02
Dong Tin	AGR	0.00	0.01		LS-2	0.75	0.70
	LS-2	0.75	0.80				

Back Propagation Neural Network Model

A three layered back propagation neural network (BPNN) modeling tool is developed using the Fortran computer language, which can run in Windows operating environments with a compatible personnel computer with a Pentium III Intel celeron processor. It allows the creation of small-scale NNs with one hidden layer using Back Propagation Neural Networks (Rumelhart et al., 1986). In back propagation neural network (BPNN) all neurons are of the sigmoid-type transfer function shown in Equation 1, and a direct input-output link is made (Solomatine and Torres, 1996).

$$Output = \frac{1}{1+\exp(-I)} \tag{1}$$

where I is a given input to a node of a corresponding output node. The BPNN describes the iterative procedure to adjust the weight by computing the mean square error (MSE) between known output value (KOV) and predicted output value (POV) response to the kth input exemplar at the end of each iteration, and propagating the error backward into the network. Before training, the weight is initialized using a random number generator ranging from –0.4 to +0.4. The MSE is calculated using following relationship;

$$MSE = \frac{1}{2}\left(KOV(k) - POV(k)\right)^2 \tag{2}$$

The training set is placed into the input data file containing four sets of values. Data for training and verification are from January 2001 to December 2001. The number of nodes in the input layer is equal to the number of pollutant indicators measured at given sampling sites in the county, for example in three layered (3-6-1) BPNN model (Figure 2), the number of inputs are considered to be three including COD, TN and TP showing three nodes on the input layers. Similarly, the number of nodes in output layers is equal to number of targets, which is one, the potential non-point pollution source. Generally a few hidden layer nodes may result in a better opportunity to capture the intricate relationship between pollutant indicators and pollutant sources as described in Lingireddy and Ormsbee, (1998).

In keeping with the above experience, in the present study hidden layer nodes are employed twice as the number of input and output target performance of the problem solution. During training, the sigmoid function computes the output always between 0 and 1. For the model prediction, a certain number is assigned for source identification as: 0 for agriculture (AGR), 0.25 for domestic (DOM), 0.5 for animal livestock (LS-1), 0.75 for aquatic livestock (LS-2) and 1.0 for the combined effect of all corresponding sources.

Calculation of SME and sources identification

In a three layered NN model, the number of nodes in the input layer depend upon three measured concentrations of pollutant indicators (N, P) in the surface runoff from land use features of the 13 sub-counties. The number of nodes in the output layer is one. A set of 208 data points is prepared from subsequent land uses (AGI, DOM, LS-1, LS-2) for training and model prediction. A total of 166 data points out of 208 are sorted randomly for the input data set to train the model, and the percentage of each land use from subsequent counties is the output layer. Further, this trained model is used for source identification. For this it uses the rest of the data points as verified data. The training process is completed using 1000 error back propagation iterations. In order to make the result more precise, 500 epochs are selected for each iteration. The model is trained from the data set of 12 sub-counties and asked for 13th. This is done by holding one data set (e.g. for Donghei Tang, 26 data

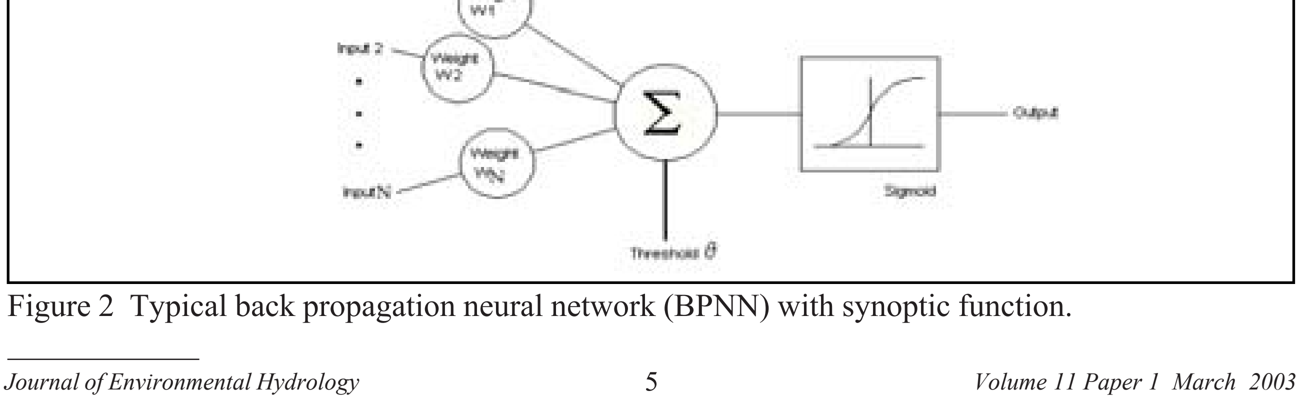

Figure 2 Typical back propagation neural network (BPNN) with synoptic function.

points), and the rest of 182 data points are used for training. The trained model is used for source identification of the Donghei Tang sub-county on all 26 data points. The same process is repeated for source identification for the rest of the counties one by one. Identified output and model identification is shown in Table 2. Many iterations are required to obtain a precise convergence of pollution source identification towards their expectations without reaching overexertion after calculation of the SME (Figure 3). In the present work, an experimental approach is used to verify the output using the input variables, which is referred to as an inverse modeling system to get the solution while using the results. The response of model is fully dependent on the independent and dependent variables.

RESULTS AND DISCUSSION

An ANN has been used as a watershed management tool for the control of environmental pollution from non-point sources. This study shows an inverse modeling approach using a BPNN for the identification of pollution sources from pollutant indicators in surface runoff released from various land uses of watershed, rather then a prediction or forecast of water quality indicators. As a sigmoid function gives the model output between 0 and 1, the values in this range have identified pollutant sources in each sub-county of the Xishan watershed. If the model prediction ranges between 0.0 and 0.24, the identified sources are AGR, otherwise they are DOM (0.25-0.49). Similarly, if the values are ranging between 0.50 and 0.74, the identified sources are LS-1, otherwise they are LS-2 (0.74 to 1.0). Overall output results show the small difference between identified output and source identification, which is because of small size of the data set. A well trained network is only expected to predict an outcome close to the actual value, but not the exact value (Figure 4).

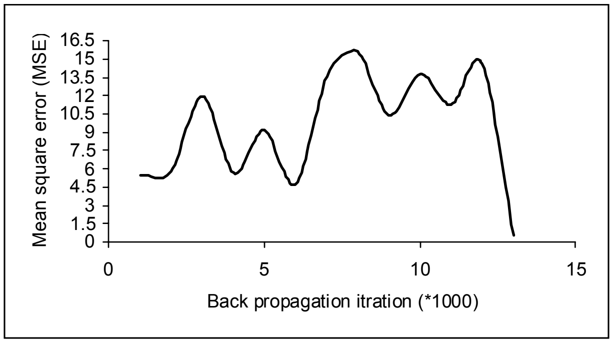

Figure 3. BPNN model training and convergence.

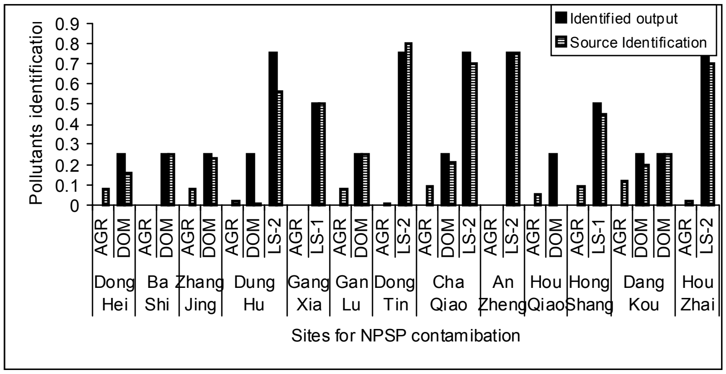

Figure 4. NNs model prediction of sources identification from all sub-counties having land use features.

Overall, the objective of the study is to predict the nature of contaminated sources qualitatively rather then quantitatively. Results are described in the given range of values for each type of contaminated source to support explicit preprocessing of the neural network,. providing a confused result for model prediction. Where the result falls between 0.00 to 0.24, the predicted contamination source is assigned to agriculture, and so on for other sources. Further verification of model prediction required a larger set of data points, which are not yet available.

The potential exists for identified sources in this Xishan watershed study to provide a auxiliary support to provide the input data profile of NPSP in the PLOAD extension of the BASIN 3.0 watershed management tool in the sub-counties. This study has the merits of identifying sources of contamination in a complex watershed having varying features of land use based on the observed pollutant indicator data. The limitation of this model is explained by changes in scenario of annual precipitation, animal unit density, average annual surface runoff, and mean annual stream flow, which need future work.

These modeling techniques can provide the management tool to handle event based risk hazards, which can be used for the planning and management of a watershed for strategic purposes. The study shows the substantial applicability of NNs for the identification of sources in storm runoff using multiple input parameters.

REFERENCES

Addiscott, T.M., A.P. Whitemore, and D.S. Powlson (1992). Farming fertilizers and the nitrate problem, CAB International, Wallingford.

Arnold, J.G., P.M. Allen and G. Bernhardt; (1993). A comprehensive ground water flow model, J. of Hydrology. 142: 47-69

Brion, G.M. and S. Lingireddy; (2000). Artificial neural networks in hydrology. Kluwer Academic Publishing, 179-197.

Chen, H. and Y. Li; (1998). Planning objective and thinking for Taihu lake water pollution prevention, Journal of Lake Sciences, Vol 10 supplement 59-66

Cooper, A.B and C.E. Thomsen; (1988). Nitrogen and phosphorus in the stream water from adjacent pastures, pine and native forest catchments. NZ J. Mar. Freshwater Res 22. 279-291.

Donigian, A.S. and W.C. Huber; (1990). Modeling of non-point source water quality and in urban and non-urban areas. EPA 68-03-3515, Environmental Res. Lab, USEPA Athens, GA.

Giorgini, A. and F. Zingales; (1986). Agriculture non-point sources pollution: model selection and application, Elsevier, Amsterdam.

Hopkinson, C.S. and J. J. Vallino. 1995. The relationship between man's activities in watersheds and rivers and patterns of estuarine community metabolism. Estuaries 18:598-621.

Kosko, B., Fuzzy Thinking, The New Science of Fuzzy Logic, Hyperion Books, 1994.

Lek, S., G. Maritxu and J. Giraudel; (1999). Prediction of stream nitrogen concentration from watershed features using neural networks. Wat. Res. 33(16), 3469-78

Lingireddy, S. and L.E. Ormsbee; (1998). Neural networks in optimal calibration of water distribution systems, artificial neural networks for civil engineering: advanced features and applications. (ed) American Society of Civil Engineers 277

Solomatine, D.P and A.A. Torres; (1996). Neural network approximation of a hydrodynamic model in optimizing reservoir operation. Intern. Proc. of the Second Int. Conference in Hydro informatics, Zurich, September 9-13, 1996.

Wen, C.G. and C.S. Lee; (1998). A neural network approach to multi-objective optimization for water quality management in a river basin, Wat. Resources Res. 34(3). 427-436

Young, R., C. Onstad, D. Bosch and W. Anderson; (1986). Agriculture non-point sources pollution model: A watershed analysis tool, Model documentation, Agriculture research service, US department of agriculture, Morris, USA.

Zhang, Q. and J.S. Stephen; (1997). Forecasting raw water quality parameters for North Saskatchewan River by neural network modeling, Wat. Res. 31(9), 2340-2351

Zaheer, I. and G. Cui; (2002a). Modeling interaction towards evaluation of non-point pollution sources, Proceeding of 2002 Summer Specialty Conference, American Water Resources Association (AWRA), July 1-3, Colorado, USA, 587-592

Zaheer, I. and B. Cui; (2002b). Application of artificial neural network for water environmental pollution control, sent for publication in International Journal of Lowland Technology, Japan.

Zaheer, I. and B. Cui; (2002c). Prediction of pollutants in the river system using a finite difference method, International Journal of Sediments Research Vol 17(2), pp 1-10, Beijing.

ADDRESS FOR CORRESPONDENCE
Zaheer Iqbal
College of Water Resources and Environment
Hohai University
Nanjing 210098
China

E-Mail: zaheeriqbal@hotmail.com