JOURNAL OF THE AMERICAN WATER RESOURCES ASSOCIATION
DECEMBER AMERICAN WATER RESOURCES ASSOCIATION 2006

LAKE WATER QUALITY ASSESSMENT FROM LANDSAT THEMATIC MAPPER DATA USING NEURAL NETWORK: AN APPROACH TO OPTIMAL BAND COMBINATION SELECTION[1]

K.P. Sudheer, Indrajeet Chaubey, and Vijay Garg[2]

ABSTRACT: The concern about water quality in inland water bodies such as lakes and reservoirs has been increasing. Owing to the complexity associated with field collection of water quality samples and subsequent laboratory analyses, scientists and researchers have employed remote sensing techniques for water quality information retrieval. Due to the limitations of linear regression methods, many researchers have employed the artificial neural network (ANN) technique to decorrelate satellite data in order to assess water quality. In this paper, we propose a method that establishes the output sensitivity toward changes in the individual input reflectance channels while modeling water quality from remote sensing data collected by Landsat thematic mapper (TM). From the sensitivity, a hypothesis about the importance of each band can be made and used as a guideline to select appropriate input variables (band combination) for ANN models based on the principle of parsimony for water quality retrieval. The approach is illustrated through a case study of Beaver Reservoir in Arkansas, USA. The results of the case study are highly promising and validate the input selection procedure outlined in this paper. The results indicate that this approach could significantly reduce the effort and computational time required to develop an ANN water quality model.

(KEY TERMS: artificial neural network; water quality; remote sensing; band combination.)

Sudheer, K.P., Indrajeet Chaubey, and Vijay Garg, 2006. Lake Water Quality Assessment From Landsat Thematic Mapper Data Using Neural Network: An Approach to Optimal Band Combination Selection. Journal of the American Water Resources Association (JAWRA) 42(6):1683-1695.

INTRODUCTION

Observations of chlorophyll-a (chl-a) and suspended sediment (SS) concentration (and many other parameters) provide quantitative information concerning water quality conditions of a water body. Accordingly, these observations can be used in various numerical schemes to help characterize the trophic status of an aquatic ecosystem. However, the number of available in situ measurements of water quality characteristics is usually limited, especially in spatial and temporal domains, because of the high cost of data collection and laboratory analysis (Panda *et al.,* 2004). As chl-a and SS are optically active water quality parameters, many researchers have employed the digital evaluation of remote sensing information at visible and near infrared (NIR) wavelengths to assess these water quality parameters in various water bodies (e.g., Ritchie *et al.,* 1990; Lathrop, 1992; Choubey, 1994).

Morel and Prieur (1977) classified water bodies into two types: Case I water body, or open ocean; and Case II water body, which is a coastal, estuary, or inland water body. In Case I, the major optically active constituent is chlorophyll, and the water does not have a great amount of suspended sediments. Consequently, the algorithms, though empirical, that relate sensor radiances to surface concentrations are effective, and the results are relatively good (Baruah *et al.,* 2001). However, for Case II water bodies, the relationship between the sensor radiance and the water quality parameters becomes complex due to the interaction of many components such as chlorophyll, suspended

[1] Paper No. 05086 of the *Journal of the American Water Resources Association* (JAWRA) (Copyright © 2006). **Discussions are open until June 1, 2007.**

[2] Respectively, Visiting Scientist, Associate Professor, and Ph.D. Candidate, Department of Biological and Agricultural Engineering, University of Arkansas, 203 Engineering Hall, Fayetteville, Arkansas 72701 (Current address/Chaubey: Department of Agricultural and Biological Engineering, Purdue University, West Lafayette, Indiana 47907) (E-Mail/Chaubey: ichaubey@purdue.edu).

sediments, and yellow substance, since all of them may be present in high concentrations. There is considerable scattering (even in NIR) from inland waters with high sediments (Baruah *et al.,* 2001). Accordingly, the relationship between the sensor data and the constituent concentration become highly nonlinear, and most of the models currently in use that are based on linear regression or on principal component analysis often fail to accurately simulate constituent concentration under such conditions (Lathrop, 1992). In this context, data driven models may be preferable that can discover relationships from input-output data even when the user does not have a complete physical understanding of the system.

In recent decades, the advent of increasingly efficient computing technology has provided exciting new tools for the mathematical modeling of dynamic systems. The ANN is one such tool that relates a set of predictor variables to a set of target variables. Artificial neural networks are well known, massively parallel computing models that have exhibited excellent performance in the resolution of complex problems in science and engineering. In recent years, the ANN technique, which is a data driven modeling tool, has become an increasingly popular tool for water quality modeling among researchers and practicing engineers (e.g., Keiner and Yan., 1998; Gross *et al.,* 1999; Tanaka *et al.,* 2000, Baruah *et al.,* 2001; Panda *et al.,* 2004).

Despite a plethora of studies on water quality modeling from remotely sensed data using ANN, there are still certain issues that are seldom addressed by the researchers. For instance, besides the fundamental question of defining an adequate neural topology, choosing the right set of input variables for approximating a function by a neural network still remains an unsatisfactorily resolved question. In remote sensing applications where correlated data of many bands are available, the selection of potential influencing variables becomes a challenge to the researchers. Constructing models such as ANN from data with nontrivial dynamics involves the problem of how to choose the best model from within a class of models or how to choose among the competing classes. The model selection problem involves selecting k nonzero elements (λ, the parameters of the model) in a given nonlinear model, g(x,λ). Following the principle of parsimony, the smallest network that adequately captures the relationships in the training data could be considered a good model (Morgan *et al.,* 2000; Sudheer, 2000).

It is observed that in most of the reported ANN-based models for water quality from remote sensing data, the input information was selected either by a trial-and-error procedure or arbitrarily. The trial-and-error procedure involves considerable computational time and requires the development and assessment of a number of models. When building models such as ANNs, it is natural to assume that having more information is better than having less. Instinctively, one might think that ANN would work better if more inputs are presented because the input vector contains all the vital information. However, in practice, this is not the case, particularly when multivariate models are developed using ANN, because inclusion of any spurious input may significantly increase the learning complexity and lead to reduced performance of the models.

Our focus in this paper is to propose an analytical approach to identify the appropriate combination of input variables (remote sensing band data) while developing ANN water quality models from remote sensing data. We also illustrate the impact the incorporation of spurious input variables has on the performance of ANN water quality models that use spectral reflectance. The central idea of the proposed method is that, based on the output sensitivity toward changes in the individual input reflectance channels (wavelength bands), a hypothesis about the importance of each band can be made and used as a guideline for selecting the appropriate input variables for modeling. We illustrate the proposed approach through a case study of Beaver Reservoir in Arkansas, USA.

ARTIFICIAL NEURAL NETWORK

An ANN attempts to mimic, in a very simplified way, human mental and neural structures and functions (Hsieh, 1993). It can be characterized as massively parallel interconnections of simple neurons that function as a collective system. The network topology consists of a set of nodes (neurons) connected by links and usually is organized in a number of layers. Each node in a layer receives and processes weighted input from previous layer and transmits its output to nodes in the following layer through links. Each link is assigned a weight, which is a numerical estimate of the connection strength. The weighted summation of inputs to a node is converted to an output according to a transfer function (typically a sigmoid function). Most ANNs have three layers or more: an input layer, which is used to present data to the network; an output layer, which is used to produce an appropriate response to the given input; and one or more intermediate layers, which are used to act as a collection of feature detectors (Figure 1).

The multilayer perceptron (MLP) is the most popular ANN architecture in use today (Dawson and Wilby, 1998). It assumes that the unknown function

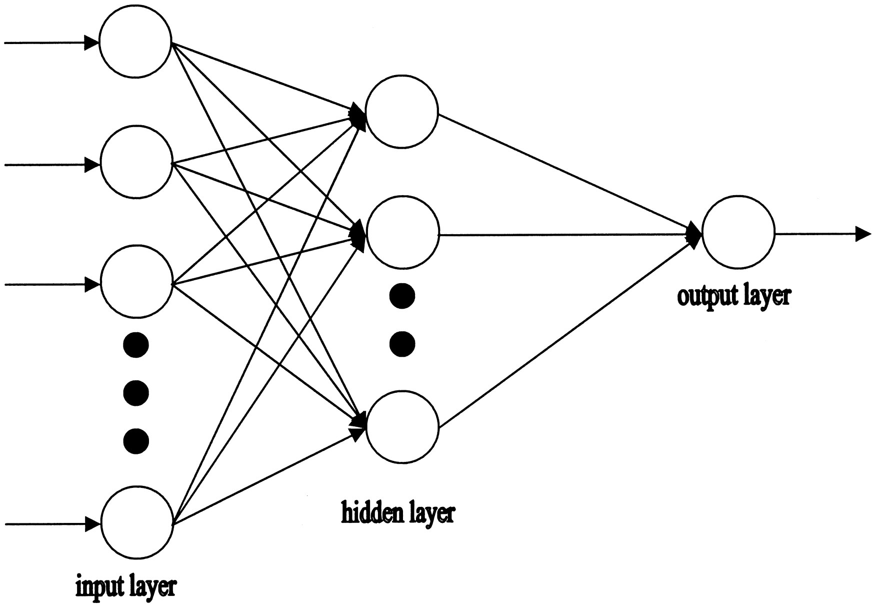

Figure 1. General Structure of a Typical Three-Layer ANN.

(between input and output) is represented by a multilayer, feed forward network of sigmoid units. The working of a three-layer ANN can be mathematically described as follows.

In an ANN model with n input neurons (x_1, …, x_n), h hidden neurons (z_1, …, z_h), and m output neurons (y_1, …, y_m), let i, j, and k be the indices representing input, hidden, and output layers, respectively. Let τ_j be the bias for neuron z_j and ϕ_k, the bias for neuron y_k. Let w_{ij} be the weight of the connection from neuron x_i to neuron z_j and β_{jk} the weight of connection from neuron z_j to y_k. The function that the ANN calculates is

$$z_j = f_A\left(\sum_{i=1}^{n} x_i w_{ij} + \tau_j\right) \quad (1)$$

$$z_j = f_A\left(\sum_{i=1}^{n} x_i w_{ij} + \tau_j\right) \quad (2)$$

where g_A and f_A are activation (transfer) functions, which are usually continuous, bounded, and nondecreasing. The most commonly employed transfer function is the logistic function, which is defined for any variable s

$$f(s) = \frac{1}{1+e^{-s}} \quad (3)$$

The training of MLP involves finding the optimal weight vector for the network. Many training techniques are available. The aim of training the network is to find a global solution to the weight matrix, which typically is a nonlinear optimization problem (White, 1989). Consequently, the theory of nonlinear optimization is applicable to the training of MLP. The suitability of a particular method is generally a compromise between computation cost and performance, and the most popular is the back propagation algorithm (Rumelhart *et al.,* 1986), which we have employed in this study.

METHODOLOGY

Absolute Variation and Residual Potential for Selection of Input

In the ANN modeling, the combination of all possible variables (input and output) locates a point in a multidimensional space (input-output space) called "phase space" (Stewart, 1989). The main tenet of the ANN methodology is that a great amount of information is contained in the sample paths in the phase space of a dynamic system beyond the usual statistics collected such as the means and variances of various output variables.

Here we consider an ANN model that represents the functional relationship y = $f(x)$ between attributes x (n-dimensional inputs) and class y that is evaluated at a set of points S (input patterns) lying inside the domain D (input range). If the magnitudes of the partial derivatives of the function with respect to the inputs are to be a measure of significance, it is implicitly assumed that the variables can change freely and independently from one another. For the analysis of data where the influencing factors can be varied individually, this assumption is valid. However, if the measured attributes are correlated (which is the case in most ANN-based models in hydrology), this assumption is not appropriate, with regard to the system represented by the ANN, because a change in one input feature may be accompanied by a change in another covariant feature.

These interrelationships could be taken into account by focusing on the variations of *f* that actually occur inside the domain D. This can be done by accounting the variation of *f* when moving between points in S. To facilitate this variation, absolute variation *v*(*f*) of the function *f*(*x*) between the points i and j could be computed and defined as the absolute value of the directional derivative of *f*(*x*) integrated along a straight line between the two points. Thus,

$$v_{ij}(f) = \int_{x_i}^{x_j} \left|\Delta f(x).u\right| dx \qquad (4)$$

where u is the unit vector in direction x_i to x_j. This variation can be ciphered in all pairs of points in S, assuming the target output values as true value of the *f*. When an attribute is insignificant to the function for the domain D, the variation in the function will be unrelated to the variation in the attribute. Thus a measure of significance of an attribute x_i for a function *f* over a dataset S would be the correlation between the absolute variation of the function $v_{ij}(f)$ and the absolute variation of that attribute $v_{ij}(x_i)$ taken between all possible pairs of points in S. Thus the variables with significant correlation between absolute variation $v_{ij}(f)$ and $v_{ij}(x_i)$ could only be taken into the input vector for ANN modeling.

However, defining the significance for an independent variable based on the correlation between the absolute variations is not trivial, as variables may possess varying degrees of correlation. In order to overcome this difficulty, we coin the term "residual potential," which we define as the difference in the correlation between absolute variations $v_{ij}(f)$ and $v_{ij}(x_i)$ and the Pearson correlation between the output and x_i. It should be noted that the correlation between absolute variations $v_{ij}(f)$ and $v_{ij}(x_i)$ is a measure of total correlation (linear and nonlinear) between the influencing variable and the output, as it accounts for the effect of perturbation of the influencing variable on the output, while the Pearson correlation gives the strength of linear relationship between the dependent and independent variables. Hence the residual potential can suggest a nonlinear correlation between the variables in question. If the relationship between x_i and the output is significantly nonlinear and sensitive, the residual potential for x_i will be positive, implying that x_i should be included in the ANN models' input vector. On the contrary, if the residual potential for x_i is negative, it implies that the relationship between x_i and the output is neither significant nor sensitive and that it need not be included in the input vector. Hence, the input vector for the ANN model can consist of only those independent variables that have a positive residual potential. We illustrate this approach for input selection through a case study.

Development of ANN Model

The development of an ANN model consists of three steps: selection of input-output variables; selection of model structure and estimation of its parameters; and validation of the identified model. In this study, we have selected the input variables according to this procedure. However, in order to validate the proposed approach to input selection, we develop a number of ANN models during the study, as we will further describe below, after first describing the general procedure we adopted for developing these ANN models.

We developed all the models by using a standard back propagation algorithm with adaptive learning and momentum rates (Nayak *et al.,* 2005) for estimating the weights. We identified the number of hidden neurons in the network that were responsible for capturing the dynamic and complex relationships among input and output variables by various trials, as no guideline currently is available to optimize it. The trial-and-error procedure started with two hidden neurons initially, and the number of hidden neurons was increased to 10 during the trials with a step size of one in each trial. For each set of hidden neurons, the network was trained in batch mode to minimize the mean square error at the output layer. In order to check any overfitting during training, we performed a cross validation by keeping track of the efficiency of the fitted model. The training was stopped when there was no significant improvement in the efficiency, and the model was then tested for its generalization properties. We selected for validation the parsimonious structure that resulted in minimum error and maximum efficiency during training as well as testing.

While developing an ANN model, generally the total available examples are divided into training and validation sets prior to the model building, and in some cases a cross validation set is also used. In most ANN applications in water resources, the data are divided arbitrarily into the required subsets. However, recent studies have shown that the way the data are divided can have a significant impact on the generalization properties of the model (Tokar and Johnson, 1999). In the present study, we have employed a method proposed by Sudheer and Jain (2004) for data division into a training set and a validation set. Their method of data division ensures representative samples from all ranges of data.

Since the sigmoid function is used as the activation function, the model output and input were scaled appropriately to fall within the function limits (0 to 1) to get over the "saturation" in training. The scaling has been performed using the maximum value of the output variable in the dataset. The convergence of the training process has been controlled by the sum of squared error (SSE) between the network output and target output.

STUDY AREA AND DATA

Beaver Reservoir

Beaver Reservoir is the principal source of drinking water to more than 300,000 people in northwestern Arkansas (Figure 2). It has a surface area of 103 km^2, a mean depth of 18 m, and a maximum depth of 60 m. The average hydraulic retention time of the reservoir is about 1.5 years. The contributing watershed area to the reservoir is 4,300 km^2. The reservoir was constructed in 1963 on the White River in northwestern

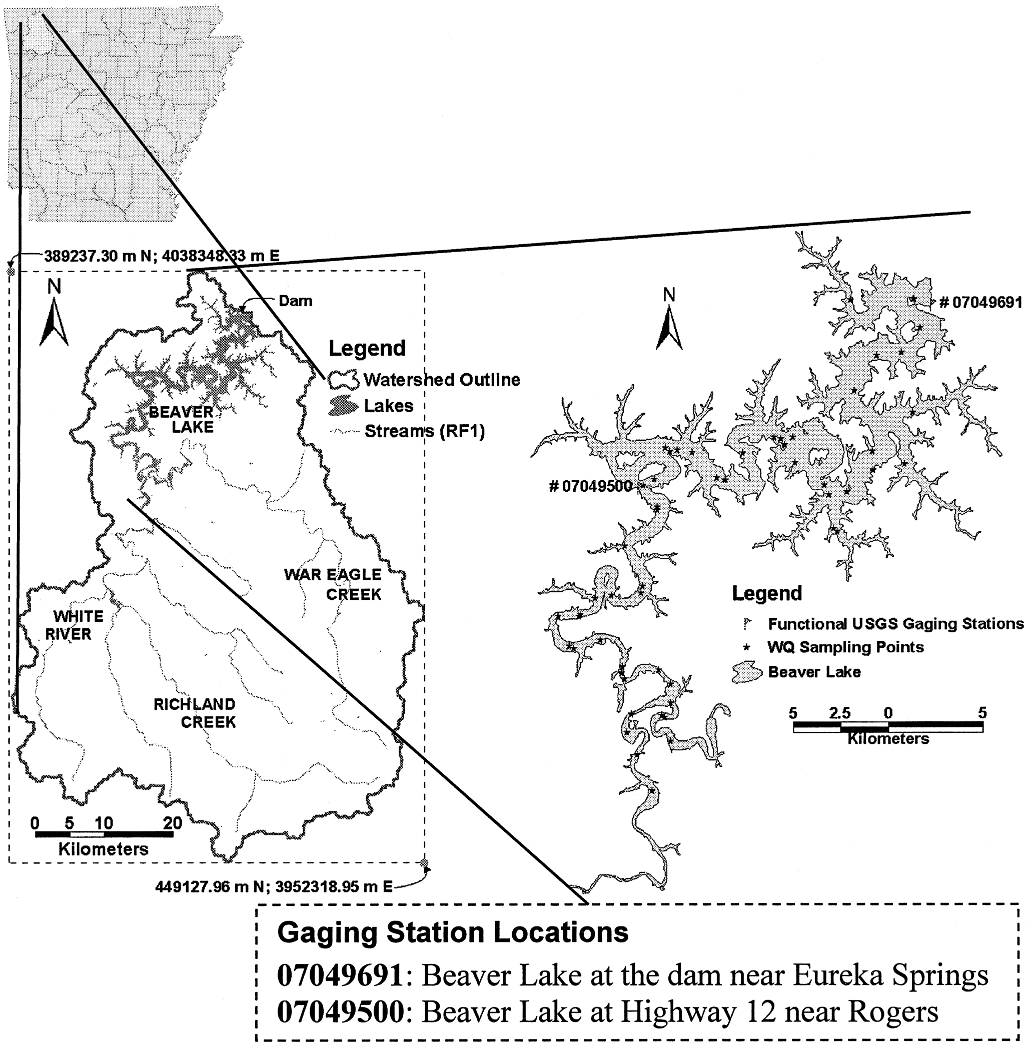

Figure 2. Location of the Study Area, Watershed Stream Network, USGS Gaging Stations, and Other Water Sample Collection Stations in Beaver Reservoir in Northwest Arkansas.

Arkansas and is operated by the U.S. Army Corps of Engineers for the purposes of hydroelectric power, flood control, and recreational activities. White River, Richland Creek, Brush Creek, and War Eagle Creek form primary tributaries to the reservoir. The major land uses within the watershed are forest (60 percent), agriculture and pasture (40 percent), and urban (2 percent). Haggard *et al.* (2003) have reported a mean total phosphorus (TP) load of 75 Mg/year into Beaver Reservoir between 1993 and 1995, with approximately 62 percent of the total load coming from the White River. The soluble reactive phosphorus (SRP) constitutes approximately 65 percent of the TP entering the reservoir and is considered bioavailable. Increased concerns over the eutrophication rate in Beaver Reservoir due to increasing urbanization and expanding agricultural production within the watershed have accelerated many investigations within this reservoir and its contributing watershed.

In Situ Water Quality Data

Water samples were collected from various spatial locations in the Beaver Reservoir on dates coinciding with TM acquisition dates (Table 1) of the reservoir for 2003 and 2004. In addition, for2001 and 2002, the chl-a data were obtained from the U.S. Geological Survey (USGS) for two of its gauging stations, No. 07049500 and No. 07049691 (Figure 2).

TABLE 1. Details of Landsat TM Data and In Situ Water Quality Sampling Dates.

Year	Water Sampling Date	Image Acquisition Date	Path/Row	Number of Water Sampling Points
2001	April 18	April 17	25/35	2
	June 13	June 13	26/35	2
	October 16	October 19	25/35	2
2002	July 10	July 9	26/35	2
	July 24	July 21	26/35	2
2003	July 21	July 21	25/35	10
	August 6	August 7	25/35	12
	December 19	December 19	26/35	12
2004	February 21	February 21	26/35	11
	April 4	April 2	25/35	11

Water sampling was done at three depths (surface, 1 m below surface, and 2 m below surface) for each location. Four liters of water samples were collected each time and stored in dark, and on ice for laboratory analyses. These samples were analyzed at University of Arkansas laboratory using standard procedures to determine SS and chl-a concentration. Average values of chl-a and SS concentrations were used for the modeling study.

Remote Sensing Data

Landsat thematic mapper (TM) data for 10 cloud free dates were acquired for the study. The TM images were precision corrected by radiometric and geometric means. The data pertaining to spectral bands were used to retrieve water quality characteristics of the Beaver Reservoir. Because the spatial resolution of band 6 (57 m) was not consistent with that of other bands (28.5 m), the radiance data for Band 6 were not considered in this study.

Beaver Reservoir has a fairly typical dendrite shape, and many samples were taken from narrow locations (Figure 2). Therefore, while extracting the radiance information from Landsat images, a single pixel gray value (digital numbers, DN) was used. The DN values for each TM band for the water sampling location on each date were extracted, after radiometric calibration, atmospheric correction, and radiometric rectification, and employed in the current study. A statistical analysis was performed on the extracted DN data to check for inconsistencies. One outlier datum was observed based on student t-test and was not considered for further analysis. The software employed for image processing in the current study was Geomatica 9.1 (PCI geomatics, Richmond Hill, Ontario, Canada) and IDRISI 32.2 (IDRISI Production, Worcester, Massachusetts).

In remote sensing, in order to improve the predictive characteristics of various band data, it is common to use various derivative indices (Thiam and Eastman, 1999; Yang and Anderson, 2000). These indices are normally derived by arithmetic manipulations of different combinations of TM band data. To evaluate the impact of such indices in ANN models for water quality retrieval, we used various indices suggested by Panda *et al.* (2004) (Table 2).

RESULTS AND DISCUSSION

Band (Input) Selection

The Pearson correlation matrix (R) between various TM bands (DNs) and the two water quality parameters (SS and chl-a) are presented in Table 3. The correlation of TM bands with chl-a range between 0.14 (for TM1) and 0.31 (for TM3), and therefore it is obvious that a linear regression between the TM data and an attribute such as chl-a would not produce much accuracy. It should be noted that four of the bands (TM2, TM3, TM4, and TM5) have comparable strength of relationship with chl-a (R = 0.27-0.31). However, in the case of SS, the strength of relationship is less for each band when compared to chl-a; the R in this case ranges from -0.20 (for TM1) to 0.30 (for TM3). The strength of linear relationship between the TM bands and SS is comparable for all the bands except TM3. A high correlation is observed between TM bands themselves (0.48 to 0.94), and it probably means that these bands are measuring similar aquatic properties in the study and that the DN values from these bands are covariant in nature. Hence a prioritization of bands for input vector is not a trivial task.

TABLE 2. Details of Various Derivative Indices Considered in the Study.

Index	Band Combination and Form
IND1	(TM1+TM2+TM3)/3
IND2	TM1/TM2
IND3	TM1/TM3
IND4	TM2/TM3
IND5	(TM5+TM7)/2
IND6	(TM4+TM5+TM7)/3
IND7	(TM1-TM3)/(TM1+TM3)
IND8	(TM1-TM2)/(TM1+TM2)
IND9	(TM2-TM3)/(TM2+TM3)
IND10	TM2+TM3
IND11	TM4+TM5+TM7

The derivative indices, in contrast to the individual band DNs, are more correlated to the water quality parameters, as is evident from the Pearson correlation matrix presented in Table 4. The sign of the correlation may be related to the magnitude of the DN values in individual bands and the arithmetic operation but can be ignored since the interest here is on the strength of relationship between the variables. It appears that a linear combination of first three bands (IND1 with R = -0.01, Table 4) may not result in an accurate model for SS when linearly regressed. However, other combinations of these bands (as in IND2, IND3, IND4, IND7, IND8, IND9, and IND10) have good potential for retrieving the SS information (R = 0.21 to -0.69, Table 4). The results also show the indices IND5, IND6, and IND11, which are derived from TM4, TM5, and TM7 (Table 2), have relatively less potential to model SS (R = 0.07 to -0.14), but have significant relationships with chl-a (R = 0.25 to 0.41). In the case of chl-a, all the indices show good correlation. Because the indices' degrees of correlation vary, a prioritization of indices is a difficult task for building ANN models.

The correlation matrix of absolute variation, computed using Equation (4), between the individual bands and the two water quality parameters is presented in Table 5. It is evident that the absolute variation provides a better picture of the strength of relationship among the variables, that is, the values in Table 5 compared with their counterparts in Table 3. A significant observation is that the data for bands TM5 and TM7 are not sensitive to SS, and hence incorporation of DNs from these bands in modeling SS would only increase the model complexity. On the other hand, TM5 and TM7 have significant relationships to chl-a and can be potential input variables in modeling. It appears that TM2 and TM3 are the dominant bands in modeling both SS and chl-a. TM1 is more sensitive to chl-a (R = 0.24) than to SS (R = 0.19) (Table 5). Furthermore, TM4 is more correlated

TABLE 3. Correlation (R) Matrix of Landsat TM DNs With Water Quality Parameters.

	TM1	TM2	TM3	TM4	TM5	TM7	SS	CHL-a
TM1	1.00							
TM2	0.92	1.00						
TM3	0.81	0.94	1.00					
TM4	0.73	0.70	0.62	1.00				
TM5	0.60	0.56	0.57	0.77	1.00			
TM7	0.51	0.48	0.49	0.70	0.94	1.00		
SS	-0.20	0.12	0.30	-0.14	-0.12	-0.11	1.00	
Chl-a	0.14	0.30	0.31	0.27	0.28	0.20	0.54	1.00

TABLE 4. Correlation (R) Matrix of Derivative Indices With Water Quality Parameters.

	IND1	IND2	IND3	IND4	IND5	IND6	IND7	IND8	IND9	IND10	IND11	SS	Chl-a
IND1	1.00												
IND2	-0.45	1.00											
IND3	-0.53	0.79	1.00										
IND4	-0.40	0.29	0.81	1.00									
IND5	0.58	-0.17	-0.35	-0.36	1.00								
IND6	0.69	-0.22	-0.35	-0.29	0.95	1.00							
IND7	-0.53	0.79	0.98	0.80	-0.31	-0.31	1.00						
IND8	-0.46	0.99	0.81	0.33	-0.16	-0.21	0.83	1.00					
IND9	-0.43	0.35	0.84	0.99	-0.38	-0.33	0.84	0.39	1.00				
IND10	0.97	-0.63	-0.71	-0.52	0.55	0.64	-0.73	-0.65	-0.56	1.00			
IND11	0.72	-0.37	-0.49	-0.37	0.90	0.97	-0.46	-0.37	-0.41	0.71	1.00		
SS	-0.01	-0.67	-0.66	-0.42	-0.12	-0.14	-0.69	-0.69	-0.46	0.21	0.07	1.00	
Chl-a	0.22	-0.55	-0.49	-0.22	0.25	0.28	-0.43	-0.52	-0.23	0.31	0.41	0.54	1.00

TABLE 5. Correlation (R) Matrix of Absolute Variance of TM Bands and Water Quality Parameter.

	TM1	TM2	TM3	TM4	TM5	TM7	SS	Chl-a
TM1	1.00							
TM2	0.88	1.00						
TM3	0.74	0.90	1.00					
TM4	0.54	0.58	0.54	1.00				
TM5	0.45	0.39	0.43	0.71	1.00			
TM7	0.34	0.29	0.30	0.69	0.93	1.00		
SS	0.19	0.50	0.57	0.22	0.00	-0.04	1.00	
Chl-a	0.24	0.33	0.30	0.19	0.23	0.17	0.55	1.00

(R = 0.22) to SS than is TM1 (R = 0.19). Again, it is evident that an assessment of the potential influencing variables based on their correlations to absolute variance is difficult.

The correlation matrix between the absolute variance of derivative indices and the two water quality parameters, presented in Table 6, suggests that a combination of TM1, TM2, and TM3 is significant in retrieving SS information from remote sensing data (R ranges from -0.69 to 0.55). It is observed that the IND5 and IND6 are not sensitive to SS as is evident from low value of correlation (R = -0.01 and 0.09, respectively). However, these indices (IND5 and IND6) show good correlation with chl-a (0.21 and 0.22). In the case of chl-a, all the indices show comparable correlation (R = -0.18 to -0.37). Similarly to the correlation of absolute variation for band data, a decision on the potential influencing derivative indices is not trivial because of the varying degrees of correlation in this study.

The residual potential, as defined in the Methodology section above, computed for each of the influencing variables is presented in Table 7. It can be observed from Table 7 that the maximum residual potential is for TM1 among the six bands for modeling both chl-a (0.10) and SS (0.39). It should be noted that TM3, TM4, TM5, and TM7 show negative residual potential in modeling chl-a, while they possess positive residual potential to model SS, though to a lesser magnitude. In the case of derivative indices, the maximum potential is observed for IND1 (0.40) to model SS, which is obvious, as this index is derived from the first three bands. A similar argument holds for IND10, with a residual potential of 0.34. All the indices derived from TM4, TM5, TM7 (IND5, IND6, and IND11) show negative residual potential to model chl-a, implying that these bands are not significant in modeling chl-a. The indices IND2 and IND8, derived from TM1 and TM2, possess a negative residual potential to model SS, suggesting that a forced nonlinear combination of these bands may not perform well

TABLE 6. Correlation Matrix of Absolute Variance of Derivative Indices With Water Quality Parameters.

	IND1	IND2	IND3	IND4	IND5	IND6	IND7	IND8	IND9	IND10	IND11	SS	Chl-a
IND1	1.00												
IND2	-0.53	1.00											
IND3	-0.62	0.61	1.00										
IND4	-0.44	0.12	0.86	1.00									
IND5	0.42	-0.13	-0.33	-0.29	1.00								
IND6	0.52	-0.23	-0.38	-0.28	0.95	1.00							
IND7	-0.65	0.66	0.97	0.80	-0.27	-0.34	1.00						
IND8	-0.55	0.99	0.64	0.17	-0.11	-0.22	0.71	1.00					
IND9	-0.49	0.18	0.88	0.99	-0.31	-0.32	0.85	0.24	1.00				
IND10	0.96	-0.68	-0.77	-0.54	0.39	0.50	-0.82	-0.71	-0.60	1.00			
IND11	0.61	-0.37	-0.50	-0.34	0.88	0.97	-0.47	-0.37	-0.39	0.61	1.00		
SS	0.39	-0.69	-0.63	-0.36	-0.01	0.09	-0.68	-0.73	-0.40	0.55	0.29	1.00	
Chl-a	0.29	-0.37	-0.36	-0.19	0.21	0.22	-0.30	-0.35	-0.18	0.32	0.32	0.55	1.00

in modeling SS. In general, the results suggest that chl-a could be modeled using TM1 and TM2 and any indices derived from these bands. For SS, all the TM bands could be employed; however, the first four are relatively significant.

TABLE 7. The Computed Residual Potential for Each Variable Corresponding SS and Chl-Concentrations.

Variable	SS	Chl-a
TM1	0.39	0.10
TM2	0.38	0.03
TM3	0.27	-0.01
TM4	0.36	-0.08
TM5	0.12	-0.05
TM7	0.07	-0.03
IND1	0.40	0.07
IND2	-0.02	0.18
IND3	0.03	0.13
IND4	0.06	0.03
IND5	0.11	-0.04
IND6	0.23	-0.06
IND7	0.01	0.13
IND8	-0.04	0.17
IND9	0.06	0.05
IND10	0.34	0.01
IND11	0.22	-0.09

Therefore, according to the proposed procedure for input selection, TM1, TM2, TM3, and TM4 could be the best combination of input vectors to model SS, while for chl-a the most influencing combination could be TM1 with TM2. The indices might not be introduced in the input vector, since those indices with positive residual potential are derived from these bands themselves. Following the principle of parsimony, they result in only model complexity and do not possess any extra information. However, the effectiveness of this approach needs to be reinforced by developing and comparing ANN models with different input combinations. Accordingly, we have developed a number of ANN models with the input variables as described in Table 8. The selection of input combinations for each model developed is done in such a way that the effect of variables that result in negative residual potential can be evaluated. Also, it is intended to evaluate the impact of derivative indices in input vector on the models' performance.

Performance of ANN Chl-a Models

The performance of the ANN models in chl-a prediction (M1 through M10) is also presented in Table 8 in terms of the common statistical indices used for performance evaluation of models (efficiency and root mean square error, RMSE). The efficiency is a measure of the models' ability to predict values away from the mean, while the RMSE indicates the residual variance. It is observed from Table 8 that the highest efficiency is produced by M3, which is made from TM1 and TM2 band information, and this confirms the earlier considerations of input selection procedure. When TM3 band data are added to the M3 model, resulting in M4, the performance slightly deteriorated, suggesting the impact of negative residual potential; TM3

TABLE 8. Input Variables of ANN Models for SS and Chl-*a* and the Corresponding Model Performance.

Model Acronym	Input Band/Index Combination	Output	Training Efficiency	Training RMSE	Validation Efficiency	Validation RMSE
M1	TM1	chl-a	30.42	15.23	30.58	15.24
M2	TM2	chl-a	32.65	14.98	32.88	14.99
M3	**TM1,TM2**	**chl-a**	**54.29**	**12.34**	**54.53**	**12.34**
M4	TM1, TM2, TM3	chl-a	54.19	12.35	54.41	12.35
M5	TM1, TM4	chl-a	5.64	17.73	5.41	17.79
M6	TM1, TM5	chl-a	5.60	17.73	5.35	17.80
M7	TM1, TM2, TM3, TM4, TM5, TM7	chl-a	53.53	12.44	53.69	12.45
M8	IND2, IND3	chl-a	47.98	13.17	48.13	13.18
M9	TM1, TM2, IND11	chl-a	50.76	12.81	50.97	12.81
M10	TM2, TM3, IND11	chl-a	15.58	16.77	15.18	16.85
M11	TM1	SS	81.55	6.50	82.36	6.35
M12	TM1, TM2	SS	92.76	4.07	92.72	4.08
M13	TM1, TM2, TM4	SS	95.82	3.09	95.78	3.11
M14	TM1, TM2, TM3	SS	95.40	3.24	95.43	3.23
M15	**TM1, TM2, TM3, TM4**	**SS**	**98.21**	**2.02**	**98.04**	**2.12**
M16	TM1, TM2, TM3, TM4, TM5	SS	88.80	5.06	88.79	5.06
M17	TM1, TM2, TM3, TM4, TM5, TM7	SS	83.24	6.19	83.28	6.18
M18	IND7, IND8	SS	85.83	5.69	85.72	5.71

Note: Efficiency is in percentage; RMSE for chl-a is in µg/l and RMSE for SS is in mg/l.

showed a negative residual potential of -0.01. The results imply that the magnitude of the residual potential is also important, as a value of -0.01 is very near to zero and may not deteriorate the models' performance if included in the input vector. This observation can be further confirmed from the performance of M5 and M6 models that a higher negative residual potential (-0.08 for TM4 and -0.05 for TM5) diminishes the performance of models built using these data (a reduction in efficiency from 30.42 percent for M1 to 5.64 percent for M5 and 5.60 percent for M6). Even though model M7 performs similarly to model M3, its input vector is not parsimonious. The inclusion of bands TM4, TM5, and TM7 in M3 (resulting in model M7) has reduced the performance of the model compared to M3. This indicates that the bands TM4, TM5, and TM7 do not contain much information to model chl-a from remote sensing data. This may be due to the absorption of light by water in these bands (TM4, TM5, and TM7), and hence no significant information could be obtained from them (Dekker and Peters, 1993). The performance of model M8, which uses IND2 and IND3, indicates that an ANN model built using derivative indices may not perform as well as a model that is developed using direct individual band information (M3). The inclusion of IND11 in model M3, resulting in M10, also confirms the earlier considerations that including a variable in the input vector that has a negative residual potential would diminish the model's performance. Hence, those variables can be considered as spurious and should not be included in the input vector. The measured and computed chl-a during training and validation for M3 is presented in Figure 3, which indicates good model performance. It is noted that the model underpredicts higher chl-a concentration during validation, probably because fewer example data were available for learning.

Performance of ANN Sediment Models

The performance of the ANN models in predicting sediment concentration in the reservoir is presented in Table 8. It is observed from Table 8 that model M15, which considers all the variables having significant positive residual potential (TM1, TM2, TM3, and TM4), performs the best among all ANN SS models. The progressive performance of the models from M11 to M15 (an increase in efficiency from 81.55 percent to 98.21 percent) reinforces the considerations of incorporating variables with positive residual potential in the input vector. The performance of M13 and M14 are comparable. The slightly better performance of M13 over M14 (a gain in efficiency of 0.42 percent)

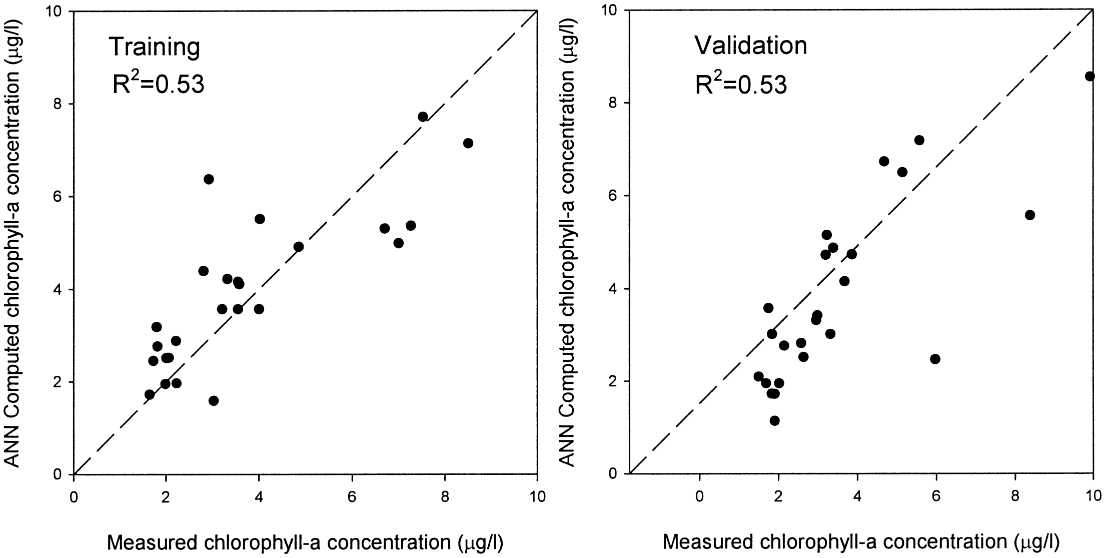

Figure 3. Scatter Plot of Measured and Computed Chl-*a* Concentrations During Training and Validation of ANN Model (M3).

confirms the effect of magnitude of the residual potential on model performance; TM4 has a higher residual potential (0.36) than TM3 (0.27). However, the performance of M16, which has the TM5 band in addition to those in M15, is relatively inferior to that of M15, even though the TM5 has a positive residual potential. This is observed with M17 also. The magnitude of the residual potential for both TM5 and TM7 in SS modeling is much less compared to the other four bands. Also, these models increase the complexity of ANN architecture, which may result in poor training due to a relatively large number of parameters to be estimated from the same number of example data. The poor performance of these models (M16 and M17) can therefore be attributed to the uncertainty in parameter estimation. Similarly to ANN chl-a models, the ANN SS model M18 built using derived indices (IND7 and IND8) does not perform well. This suggests that the nonlinear relationship between the band data and the SS can be captured well by the ANN model directly, even without inducing any transformation. On the contrary, derivative indices lose some nonlinear information when transformed, which may be the reason for an inferior performance of M18 relative to M11.

The sediment concentrations computed by M15 are presented in a scatter plot against their measured counterparts in Figure 4. The scatter is close to the 45-degree line and indicates a good performance during training as well as validation. The training and validation datasets contained all ranges of data and hence show good generalization properties for the model. On the other hand, a model built by arbitrary division of data into training and validation datasets did not perform well during validation (for brevity the results are not presented here), suggesting the importance of data division while ANN model building.

SUMMARY AND CONCLUSIONS

In this study we propose a method that establishes the output sensitivity toward changes in the individual input reflectance channels while modeling water quality from remote sensing data. From the sensitivity, a hypothesis about the importance of each band can be made and used as a guideline to select appropriate input variables (band combination) for developing ANN models for water quality retrieval. The approach is illustrated through a case study of Beaver Reservoir. The results of the case study validate the input selection procedure outlined herein. It is observed that a linear regression-based modeling for water quality information retrieval from remote sensing data may not result in good results, as the relationship between the radiance and water quality parameters is highly nonlinear. While this observation has been reported by many researchers, the analyses of the data used in this study reinforce it. The study suggest that using derivative indices as input to develop ANN models is not an appropriate approach, as the performance of models developed based on the indices did not show good performance; rather, they result in reduced model performance. However, these indices may explain more variance in the data by linear regression as opposed to using individual bands. Overall, the study demonstrates that significant

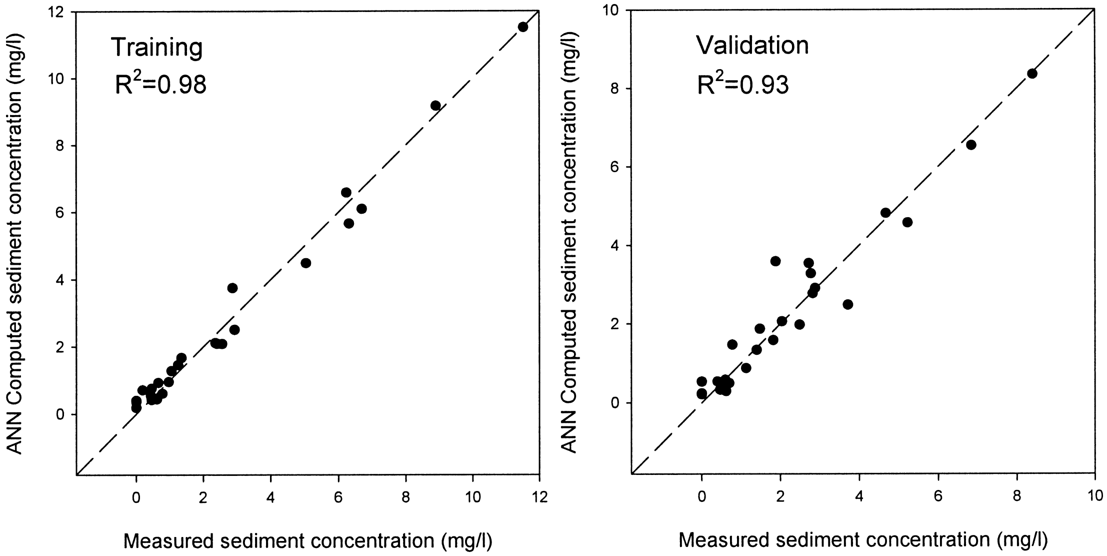

Figure 4. Scatter Plot of Measured and Computed SS Concentrations During Training and Validation of ANN Model (M15).

influencing variables for modeling a water quality parameter can be identified by analyzing the strength of relationship between absolute variance of individual variable and the targeted output and can lead to the development of parsimonious ANN models.

ACKNOWLEDGMENTS

The first author would like to acknowledge the BOYSCAST Fellowship he received from the Department of Science and Technology, Government of India. This study was funded by the USEPA through grant No. X797654601, University of Arkansas-Division of Agriculture, and Southern Regional Project S-1004: Development and Evaluation of TMDL Planning and Assessment Tools and Processes.

LITERATURE CITED

Baruah, P.J., M. Tumura, K. Oki, and H. Nishimura, 2000. Neural Network Modeling of Lake Surface Chlorophyll and Sediment Content From Landsat TM Imagery. Proceedings of the 22nd Asian Conference on Remote Sensing, Singapore.

Choubey, V.K., 1994. Monitoring Water Quality in Reservoirs With IRS-1A-LISS-I. Water Resources Management 8:121-136.

Dawson, D.W. and R. Wilby, 1998. An Artificial Neural Network Approach to Rainfall-Runoff Modeling. Hydrological Sciences Journal 43(1):47-65.

Dekker, A.G. and S.W.M. Peters, 1993. The Use of the Thematic Mapper for the Analysis of Eutrophic Lakes: A Case Study in The Netherlands. International Journal of Remote Sensing 14(5):799-822.

Gross, L., S. Thiria, and R. Frouin, 1999. Applying Artificial Neural Network Methodology to Ocean Color Remote Sensing. Ecological Modeling 120: 237-246.

Haggard, B.E., P.A. Moore, Jr., I. Chaubey, and E.H. Stanley, 2003. Nitrogen and Phosphorus Concentrations and Export From an Ozark Plateau Catchment in the United States. Biosystems Engineering 86:75-85.

Hsieh, C., 1993. Some Potential Applications of Artificial Neural Networks in Financial Management. Journal of Systems Management 44(4):12-15.

Keiner, L.E. and X.H. Yan, 1998. A Neural Network Model for Estimating Sea Surface Chlorophyll and Sediments From Thematic Mapper Imagery. Remote Sensing of Environment 66:153-165.

Lathrop, R., 1992. Landsat Thematic Mapper Monitoring of Turbid Inland Water Quality. Photgrammetric Engineering and Remote Sensing 58:465-470.

Morel, A. and L. Prieur, 1977. Analysis of Variation in Ocean Color. Limnology and Oceanography 22:709-722.

Morgan, P., B. Curry, and M. Beynon, 2000. Pruning Neural Networks by Minimization of the Estimated Variance. European Journal of Economic and Social Systems 14(1):1-16.

Nayak, P.C., K.P. Sudheer, D.M. Rangan, and K.S. Ramasastri, 2005. Short-Term Flood Forecasting With a Neurofuzzy Model. Water Resources Research 41, W04004, doi:10.1029/2004 WR003562.

Panda, S.S., V. Garg, and I. Chaubey, 2004. Artificial Neural Network Application in Lake Water Quality Estimation Using Satellite Imagery. Journal of Environmental Informatics 4(2):65-74.

Ritchie, J.C., M.C. Charles, and F.R. Schibe, 1990. The Relationship of MSS and TM Digital Data With Suspended Sediments, Chlorophyll, and Temperature in Moon Lake, Mississippi. Remote Sensing and Environment 33:137-148.

Rumelhart, D.E., G.E. Hinton, and R.J. Williams, 1986. Learning Representations by Back Propagating Errors. Nature 323:533-536.

Stewart, I., 1989. Does God Play Dice? The Mathematics of Chaos. Blackwell, Cambridge, Massachusetts.

Sudheer, K.P., 2000. Modeling Hydrological Processes Using Neural Computing Technique. Ph. D. Thesis, Indian Institute of Technology, Delhi, India.

Sudheer, K.P. and A. Jain, 2004. Explaining the Internal Behaviour of Artificial Neural Network River Flow Models. Hydrological Processes 18(4):833-844.

Tanaka, A., M. Kishino, T. Oishi, R. Doerffer, and H. Sciller, 2000. Application of the Neural Network Method to Case II Water. *In:* SPIE Proceedings, Remote Sensing of Ocean and Sea Ice 2000, Barcelona, 4172:144-152.

Thiam, S. and R.J. Eastmen, 1999. Vegetation Indices. Guide to GIS and Image Processing, Vol 2, Idrisi Production, Clarke University, Worcester, Massachusetts, Vol. 2, pp. 107-122.

Tokar, A.S. and A. Johnson, 1999. Rainfall-Runoff Modeling Using Artificial Neural Networks. Journal of Hydrologic Engineering, ASCE 4(3):232-239.

White, H., 1989. Learning in Artificial Neural Networks: A Statistical Perspective. Neural Computation 1:425-464.

Yang, C. and G.L. Anderson, 2000. Mapping Grain Sorghum Yield Variability Using Airborne Digital Videography. Precision Agriculture 2:7-23.