Prediction of weekly nitrate-N fluctuations in a small agricultural watershed in Illinois

Momcilo Markus, Mohamad I. Hejazi, Peter Bajcsy, Orazio Giustolisi and Dragan A. Savic

ABSTRACT

Agricultural nonpoint source pollution has been identified as one of the leading causes of surface water quality impairment in the United States. Such an impact is important, particularly in predominantly agricultural areas, where application of agricultural fertilizers often results in excessive nitrate levels in streams and rivers. When nitrate concentration in a public water supply reaches or exceeds drinking water standards, costly measures such as well closure or water treatment have to be considered. Thus, having accurate nitrate-N predictions is critical in making correct and timely management decisions. This study applied a set of data mining tools to predict weekly nitrate-N concentrations at a gauging station on the Sangamon River near Decatur, Illinois. The data mining tools used in this study included artificial neural networks, evolutionary polynomial regression and the naive Bayes model. The results were compared using seven forecast measures. In general, all models performed reasonably well, but not all achieved best scores in each of the measures, suggesting that a multi-tool approach is needed. In addition to improving forecast accuracy compared with previous studies, the tools described in this study demonstrated potential for application in error analysis, input selection and ranking of explanatory variables, thereby designing cost-effective monitoring networks.

Key words | artificial neural networks, drinking water, forecasting, genetic algorithms, naive Bayes model, nitrate-N

Momcilo Markus (corresponding author)
Institute of Natural Resource Sustainability,
University of Illinois at Urbana-Champaign,
2204 Griffith Dr,
Champaign, Illinois 61820,
USA
E-mail: *mmarkus@illinois.edu*

Mohamad I. Hejazi
Ven-Te Chow Hydrosystems Laboratory,
Department of Civil and Environmental Engineering,
University of Illinois at Urbana-Champaign,
205 North Mathews Ave,
Urbana, IL 61801,
USA

Peter Bajcsy
National Center for Supercomputing Applications,
University of Illinois at Urbana-Champaign,
1205 West Clark Street,
Urbana, IL 61801,
USA

Orazio Giustolisi
Department of Civil and Environmental Engineering,
Technical University of Bari,
II Engineering Faculty,
Taranto via Turismo 8, 74100,
Italy

Dragan A. Savic
Centre for Water Systems, School of Engineering, Computing and Mathematics,
University of Exeter,
Harrison Building, North Park Road,
Exeter EX4 4QF,
UK

INTRODUCTION

Many communities in the Midwestern United States have been facing frequent water quality problems related to an excessive concentration of nitrate-nitrogen (nitrate-N) in drinking water sources. The maximum contaminant level (MCL) for nitrate-N was set by the United States Environmental Protection Agency (USEPA 1991) at 10 milligrams per liter (mg/L). Water supply utilities and municipalities are required to develop plans to reduce nitrate-N concentrations below the MCL. When nitrate-N concentration in a public water supply reaches or exceeds drinking water standards, costly measures such as well closure or water treatment have to be considered. Accurately predicting such incidents of high nitrate-N concentration ahead of time is critical in water supply management. The prediction models rely on determining which important parameters control short-term fluctuations in nitrate-N

doi: 10.2166/hydro.2010.064

concentrations in water and developing a procedure that accurately predicts nitrate-N concentrations under different conditions. The traditional approach to nitrate-N prediction is typically based on deterministic physical models, calibrated for historical conditions and applied to predict future water quantity and quality. Those models require preparation of large input datasets and a time-consuming calibration and validation process. An alternative to using traditional conceptual modeling is using data mining techniques. Examples include artificial neural networks (ANN) (Maier & Dandy 1996; Markus *et al.* 2003; Sharma *et al.* 2003; Suen & Eheart 2003; Mishra *et al.* 2004; Yu *et al.* 2004), genetic algorithms (GA) (Goldberg 1989; Bobbin & Recknagel 2001; Muttil & Lee 2005), evolutionary polynomial regression (EPR) (Giustolisi & Savic 2006; Giustolisi *et al.* 2007, 2008; Doglioni *et al.* 2008) and naive Bayes methods (NBM) (Bajcsy *et al.* 2004; Peng *et al.* 2004; Kumar *et al.* 2006). These data-driven methods could capture important relationships in complex multivariate datasets that are not easily detected using traditional approaches.

This research is an extension of the 2003 Markus *et al.* study. Using the same datasets that Markus *et al.* used previously, this study applies ANN, EPR and NBM methods to fine-tune the predictions of weekly nitrate-N concentrations in the Upper Sangamon River watershed in central Illinois. The origins of nitrates, long-term trends in nitrate concentration, climate variations and changes in land use are beyond the scope of this paper. Instead, the methods applied herein use the observed weekly river discharge, air temperature, precipitation and nitrate-N concentration to predict short-term (weekly) nitrate-N fluctuations.

CASE STUDY

The Upper Sangamon River watershed, shown in Figure 1, discharges into Lake Decatur, a water supply reservoir for the City of Decatur, Illinois. The drainage area upstream of the Lake Decatur watershed is approximately 2,374 km^2. Agriculture is the dominant land use within the Upper Sangamon watershed. Row crops (corn and soybeans) cover approximately 87% of the total watershed area. Most water quality problems in the Sangamon River are associated with nonpoint source pollution generated in the Upper Sangamon River watershed. The hydrologic and meteorological data used in this study (Figure 2) were obtained from the Upper Sangamon River near Decatur, Illinois, during January 1994 to April 1999 (Keefer & Demissie 2000). Datasets included weekly average nitrate-N concentration, N_t, in units of milligrams per liter (mg/L); weekly average flow discharge, Q_t, expressed in cubic meters per second (m^3/s); weekly average temperature, T_t, in degrees Celsius (°C) and total weekly precipitation, P_t, in centimeters (cm), where t represents time in weeks. Measurements were divided into training and testing datasets. Half of the dataset was used in training and the other half was used in model testing. For the 1994–1999 time period used in this study, samples were taken each year during the high nitrate concentration season, which typically starts in April and ends in October.

INPUT SELECTION

The selection of inputs is a critical step in model building. In building ANN models, this complex task "has received little attention" (Bowden *et al.* 2005*a*). Markus (2005) recommended adopting the fully automated ANN with automatic input selection. Bowden *et al.* (2005*a*,*b*) described several input selection methods. Nonetheless, to facilitate a comparison with a previous study the inputs were adopted from Markus *et al.* (2003). The study used a trial-and-error approach with various inputs and lag times. Markus *et al.* (2003), however, determined the two sets of inputs producing maximum forecast accuracy for future weekly nitrate-N concentration, N_{t+1}. The first set included four current weekly inputs: N_t, Q_t, T_t and P_t, and the second set included seven current and previous weekly inputs: N_t, Q_t, T_t, P_t, Q_{t-1}, T_{t-1} and P_{t-1}. The four-input set has shown slightly better results and was adopted for all ANN-based models in this study. EPR models, on the other hand, have a capability to select a subset of inputs and the relationship type relevant for model predictions (Giustolisi & Savic 2006; Giustolisi *et al.* 2007, 2008; Doglioni *et al.* 2008). Although the EPR model could be presented with a large number of inputs, it selects only the relevant ones.

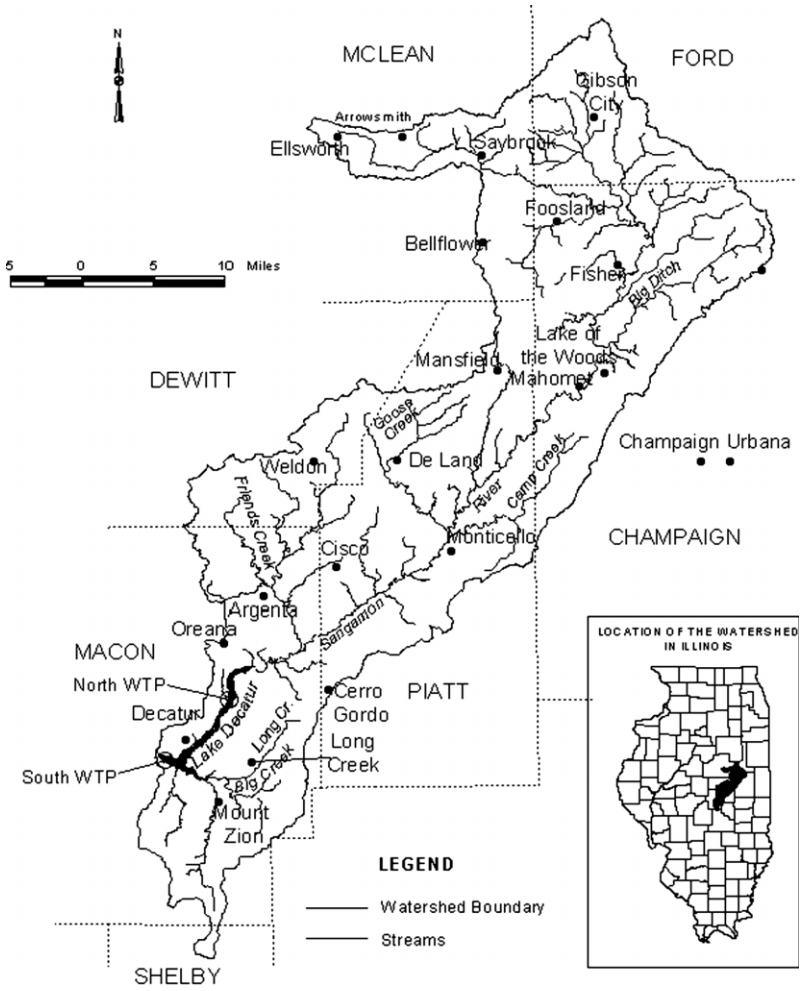

Figure 1 | The Upper Sangamon River watershed in Central Illinois.

For that reason, the seven-input set, the larger of the two, was used as an initial dataset for EPR. For the NBM, both four- and seven-input sets were used.

METHODOLOGY

The following models were applied to predict one-week-ahead nitrate-N concentration: (i) ANN back-propagation, denoted as ANN1, ANN2, ANN3 and ANN4, for one, two, three, and four hidden nodes, respectively; (ii) Evolutionary Polynomial Regression (EPR) and (iii) the naive Bayes model (NBM). All the models predicted N_{t+1} as a function of previous observations of the monitored variables N_t, Q_t, P_t and T_t. These datasets were first standardized by subtracting the mean and dividing by the standard deviation. After obtaining the model outputs in the standard domain, data were transformed back to the original domain.

Back-propagation neural network (ANN)

Artificial neural networks can be defined as a parallel interconnected network of simple elements and their hierarchal organizations (Kohonen 1988). ANN models have been applied to rainfall forecasting (French *et al.* 1992), rainfall–runoff modeling (Giustolisi & Laucelli 2005), runoff forecasting (Tokar & Markus 2000; Zhang & Govindaraju 2003; Moradkhani *et al.* 2004), water quality modeling (Maier & Dandy 1996; Bowden *et al.* 2006; Amenu *et al.* 2007; Stenemo *et al.* 2007), groundwater level prediction (Giustolisi & Simeone 2006), synthetic

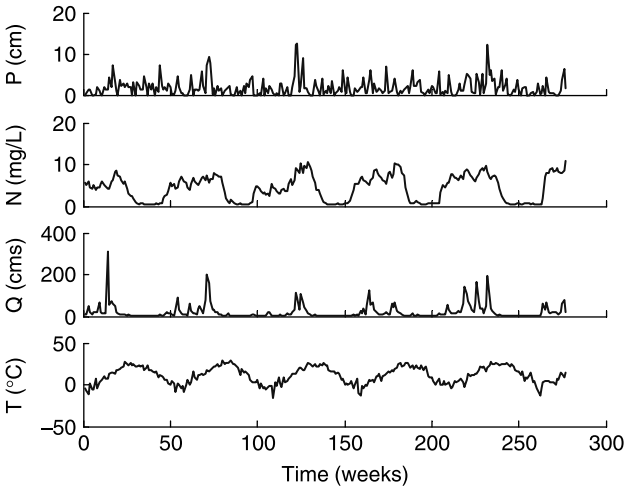

Figure 2 | The Sangamon River weekly mean data (*T*-air temperature, *Q*-discharge, *N*-nitrate-N concentration, *P*-precipitation).

data generation (Ochoa-Rivera *et al.* 2002, 2007; Markus 2006) and hydrologic classification (Hall & Minns 1999; Thandaveswara & Sajikumar 2000). Model predictions are evaluated through back-propagation, which performs computations backward through the network. A neural network consists of input, hidden and output layers. The ANN approach adopted in this study uses the Gradient Descent Back-Propagation method, in which training is accomplished by updating the model parameters (weights and biases) in the direction of the steepest negative gradient of the performance function (Salas *et al.* 2000; Markus 2006).

The ANN algorithm used in this study was based on the neural networks toolbox in MATLAB (Mathworks 2007) and used a cross-validation method. The algorithm stopped if any of the following stopping criteria was met: maximum number of epochs, minimum performance gradient or performance goal. The method also used a variable learning rate and momentum terms.

The network output y_j can be expressed as (Jain 2008)

$$y_j = f\left(\sum_{i=1}^{N} W_i X_i + b_j\right) \tag{1}$$

where W_i and b_j are network parameters, $f(\cdot)$ is an activation function and E is the network error, as follows:

$$f(x) = \frac{1}{1+e^{-x}} \tag{2}$$

$$E = \sum_{P}\sum_{N}(y_j - t_j)^2 \tag{3}$$

In Equation (3), N, P, y_j and t_j are the number of output nodes, the number of training patterns, computed output and observed output, respectively.

Evolutionary polynomial regression (EPR)

Genetic programming (GP) has gained much popularity because of its evolutionary methodology, which is used to search a symbolic mathematical expression and approximate the structural form of mathematical relationships. GP combines an efficient problem-solving procedure with powerful symbolic representations (Koza 1992). This type of problem is often called symbolic regression, and is classified as a grey-box model. Unlike neural networks, GP establishes relations that can be viewed and possibly interpreted and does not require a predefined structure. However, GP lacks the capability to optimize coefficients efficiently and grows substantially in length very quickly (Davidson *et al.* 1999, 2000). Starting from the main GP drawbacks, Giustolisi & Savic (2006) developed an evolutionary modeling approach called Evolutionary Polynomial Regression (EPR), which draws its strength from a two-stage procedure: a genetic algorithm identifies the model structures and a numerical least-squares regression estimates the coefficients in the selected expressions. The result is a set of models returned as formulae. EPR was successfully applied to environmental modeling problems by Giustolisi *et al.* (2007, 2008) and Doglioni *et al.* (2008). In Figure 3 a sketch of the EPR framework and its major components is given. Giustolisi & Savic (2006) provide full details of this method.

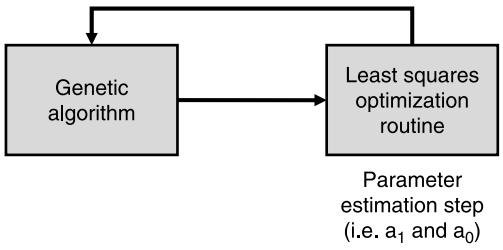

Figure 3 | Simplified approach of evolutionary polynomial regression procedure.

Naive bayes model (NBM)

The naive Bayes model is used to explore the relationships between the dependent variable and explanatory variables. Currently, interest is emerging within bioinformatics to use various kinds of Bayesian methods (Bajcsy *et al.* 2004). Newer naive Bayes inference is orders of magnitude faster than Bayesian network inference using Gibbs sampling and belief propagation. Newer methods could also be augmented using local Markov dependence among observations (Peng *et al.* 2004). Naive Bayes represents a distribution as a mixture of components, where within each component all variables are assumed independent of each other. The naive Bayes model can be used for classifying samples based on applying Bayes' theorem with "naive" independence assumptions (Bajcsy *et al.* 2006).

Bayesian techniques may provide more realistic coefficient and standard deviation estimates using less data than Gaussian techniques do. A multiple naive Bayesian model has been build for this study site as used in many similar studies (Bajcsy *et al.* 2006; Kumar *et al.* 2006). The naive Bayes model computes the posterior probability (P) for the output variable [nitrate-N at time $t+1$: N_{t+1}] conditioned by input variables (e.g. nitrate-N, phosphorus, discharge, temperature at time t) $P[N_{t+1}|N_t, P_t, Q_t, T_t]$ from the joint probability, e.g. $P[N_{t+1}, N_t, P_t, Q_t, T_t]$ over the evidence, e.g. $P[N_t, P_t, Q_t, T_t]$. Using the "naive" conditional independence assumption that each input variable is independent of every other input variable, the joint probability will be substituted by the prior probability of output variable $P[N_{t+1}]$ and the likelihood of each input variable conditioned by output variable $P[N_t, P_t, Q_t, T_t|N_{t+1}]$. By inspecting multiple conditional probabilities, conclusions can be derived about the nitrate-N levels due to an increase/decrease in input variables.

Forecast evaluation

Specific forecasts in this study were evaluated using root-mean-square error (RMSE), Nash–Sutcliffe efficiency index (NSEI) and forecast bias (B). A forecast error at time t ($t = 1, 2, \ldots, n$) can be expressed as $e_t = \hat{N}_t - N_t$, where $\hat{N}_t$ and N_t are predicted and observed nitrate-N concentrations at time t, respectively. Then, the RMSE is

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2} \tag{4}$$

The Nash–Sutcliffe efficiency index, NSEI, is expressed as

$$\mathrm{NSEI} = 1 - \frac{\sum_{t=1}^{n} e_t^2}{\sum_{t=1}^{n} (N_t - \bar{N})^2} \tag{5}$$

where $\bar{N}$ is the mean of the observed nitrate-N concentrations. NSEI ranges between 0 and 1 (perfect forecast).

Forecast bias is expressed as

$$B = \frac{1}{n}\sum_{t=1}^{n} e_t \tag{6}$$

The forecasts are also evaluated in a categorical mode for which the rationale comes from practical applications of the nitrate-N forecasting model. In their daily operations, City of Decatur water managers apply an emergency plan when the nitrate-N concentration exceeds 8.5 mg/L. Such binary categorical forecasting is illustrated in Table 1, where cases a, b, c and d are defined as counts of false positive, accurate negative, accurate positive and false negative predictions, respectively.

The False Alarm Ratio (FAR) is one of the most commonly used ratios in literature (Haklander & Van Delden 2003). The FAR is calculated as a ratio:

$$\mathrm{FAR} = \frac{a}{a+c} \tag{7}$$

Table 1 | Outcomes of binary forecasting in this study

Nitrate-N		Predicted concentration > 8.5 mg/L? Yes	No
Observed concentration >8.5 mg/L?	No	False positive (a) (false alarm)	Accurate negative (b)
	Yes	Accurate positive (c)	False negative (d)

High values of FAR are indicative of poor model performance in accurately predicting nitrate-N concentration in excess of 8.5 mg/L. FAR ranges between zero and one; if FAR = 0, no prediction is false positive, i.e. the model positive predictions (that nitrate-N concentration will exceed the threshold) are always accurate; if FAR = 1, all positive predictions are inaccurate, i.e. false alarms.

The Critical Success Index (CSI) additionally incorporates false negative counts and can be expressed as (Roebber *et al.* 2002)

$$\mathrm{CSI} = \frac{c}{a + c + d} \tag{8}$$

CSI ranges between zero (best) and one (worst). It does not account for accurate negative predictions and is often regarded as an index that considers only those situations in which a forecasting problem exists (Haklander & Van Delden 2003). This ratio appears appropriate for the high threshold of the forecasting problem in this study, which is dominated by accurate negative outcomes (b). CSI is biased, however, because it inflates warning skill with increasing event frequency (Haklander & Van Delden 2003).

The Heidke Skill Score (HSS) (Benedetti *et al.* 2005) can be expressed as

$$\mathrm{HSS} = \frac{2(cb - ad)}{a^2 + d^2 + 2cb + (a + d)(c + b)} \tag{9}$$

where HSS ranges between 0 and 1 (perfect forecast).

The categorical forecast bias (CB) (Eder *et al.* 2006) is calculated as

$$\mathrm{CB} = \frac{a + c}{c + d} \tag{10}$$

For an unbiased model, CB = 1. Departure from 1 indicates bias.

RESULTS

ANN

A batch gradient descent back-propagation algorithm with multiple nodes was used to optimize the parameters of the artificial neural network (ANN). The maximum number of epochs, the minimum performance gradient and the performance goal were 100,000, 1E-10 and zero, respectively. A cross-validation process was used as an additional stopping criterion to avoid over-fitting.

The ANN was run with four input variables (N_t, Q_t, T_t, P_t) to predict weekly nitrate-N concentration (N_{t+1}). Five different models, each having a single hidden layer, with variable numbers of hidden nodes ranging between 1 and 5 were applied (Figure 4). The models were denoted as ANN1, ANN2, ANN3, ANN4 and ANN5, with one, two, three, four and five hidden nodes, respectively. The model with two hidden nodes had the smallest testing error. Although including more than two nodes would improve the prediction accuracy of the training data, it would lead to a reduction in accuracy for the testing data. With a two-node ANN model, the minimum RMSE values of training and testing data were 0.787 mg/L and 0.935 mg/L, respectively.

EPR

Including all seven explanatory variables, EPR provided two optimal expression forms to predict weekly nitrate-N levels. With 200 generations and cross-validation, the two derived models EPR1 and EPR2 are shown in Equations (11) and (12), respectively:

$$N_{t+1} = 0.827\, N_t \tag{11}$$

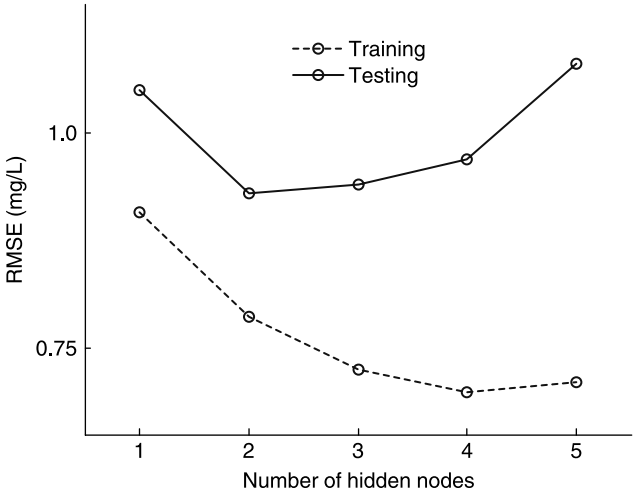

Figure 4 | RMSE as a function of the number of hidden nodes in the ANN model.

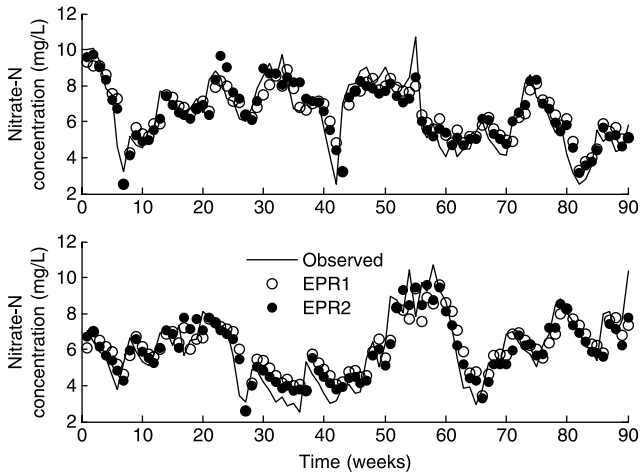

Figure 5 | Observed and simulated weekly nitrate-N concentration for training (top) and testing (bottom) using EPR modeling.

$$N_{t+1} = 0.659\,N_t + 0.560\,N_t\sqrt{Q_t} \tag{12}$$

Figure 5 shows both training (top) and testing (bottom) results for the EPR1 and EPR2 models. Testing RMSE of the EPR1 model was equal to 1.170 mg/L. The EPR2 model was somewhat more accurate with a testing RMSE of 1.010 mg/L. The difference in NSEI, however, appeared more significant. For EPR1, NSEI was 0.659, and for EPR2, NSEI was 0.742.

Unlike other commonly used data mining techniques, EPR selects relevant inputs and provides a functional model form. It is data-driven and often discovers relationships not easily acquired by other methods. Both Equations (11) and (12), for example, indicate that future nitrate-N concentration is proportional to current nitrate-N concentration, indicating that weekly nitrate-N concentration time series have a strong autocorrelation. Also, Equation (12) indicates that the future nitrate-N correlation is proportional to the current discharge, which is consistent with numerous other studies (Cohn *et al.* 1992; Guo *et al.* 2002). The product between N_t and the square root of Q_t in Equation (12) could also indicate that the correlation between these two variables is proportional to nitrate-N concentration. Indeed, during the high-nitrate season, N_t and Q_t are highly correlated, and vice versa; during the low-nitrate season the discharge peaks are less frequently accompanied by increases in nitrate concentration.

NBM

The NBM model used two categories, low and high values, for each explanatory variable. The categories were separated by the average observed value as a threshold, except for nitrate-N concentration, in which case the low and high categories were separated using the emergency cutoff level of 8.5 mg/L.

Two models were tested, $N_{t+1} = f[N_t, Q_t, P_t, T_t]$ (NBM1) and $N_{t+1} = f[N_t,\ Q_t,\ Q_{t-1},\ P_t,\ P_{t-1},\ T_t,\ T_{t-1}]$ (NBM2). The model testing results (Table 2) indicate that NBM1 accurately predicted 79 of 80 low concentrations, but only 2 of 9 high concentrations. It also exhibited some bias, as the number of predicted high flows (3) was less than the number of the observed ones (9). On the other hand, for NBM2, the number of predicted high flows (10) was similar to the number of the observed ones (9). However, NBM2 had a much larger number of false alarms (7), compared to NBM1 (1).

Naive Bayes models offer additional analyses. Figure 6 shows a conditional probability that the predicted N_{t+1} will be greater than 8.5 mg/L (herein denoted as high), given that N_t and also all other input values were high. For inputs other than nitrate-N concentration, the values above average were considered high. Consequently, nitrate-N concentration below 8.5 mg/L and other variables below mean were considered "low". Figure 6 indicates that, if all input values were high, the output will be high with a 79.0% probability. Thus, there is a 21.0% false alarm risk given that all inputs were high. For NBM2, this risk is only about 3%, but also having all seven inputs above their thresholds would be extremely rare. These conditional analyses also could provide an alternative method to ranking input variables by their importance, providing monitoring programs with valuable input. Those variables with higher

Table 2 | Naive Bayes analysis model test results. Numbers in the table denote the count of weeks for each category. Test sample size is 89 weeks. Letters in parentheses correspond to the outcomes of binary forecasting given in Table 1

Naive bayes model 1 (NBM1)				Naive bayes model 2 (NBM2)			
		Predicted				Predicted	
Data		Low	High	Data		Low	High
Observed	Low	79 (*b*)	1 (*a*)	Observed	Low	73 (*b*)	7 (*a*)
	High	7 (*d*)	2 (*c*)		High	6 (*d*)	3 (*c*)

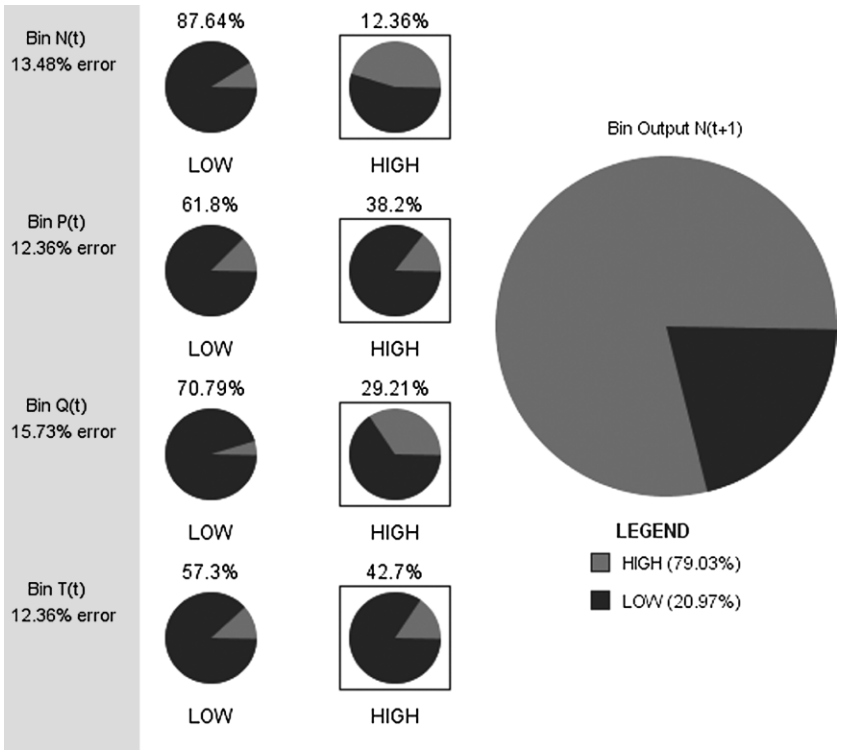

Figure 6 | Test results of naive Bayes model 1 (NBM1).

effects on the prediction accuracy should have higher importance than those with less significant effects.

SUMMARY

A comparison among all models in this study is presented in Table 3 for the training dataset and Table 4 for the testing dataset. These tables show the measures of forecasting accuracy for six ANN models, two EPR models and two NBM models. For comparison, ANN0 denotes the results of the previous study (Markus *et al.* 2003), ANN1–ANN5 denote ANN models with one through five hidden nodes; EPR1 and EPR2 are given by Equations (11) and (12); and NBM1 and NMB2 are naive Bayes models with four and seven inputs, respectively. The test results (Table 4) show that, for specific forecasting, the ANN model with two hidden nodes (ANN2) was the most accurate in terms of RMSE. The results also show that the EPR2 model was the most accurate in terms of NSEI and *B*. For categorical forecasting, the models with the best CSI and HSS were ANN1, ANN2 and ANN4; the model with the best FAR was EPR2, and the model with the best CB was the NBM2 model. While most of the forecast evaluation statistics varied considerably, RMSE and *B* were relatively constant across all the models. RMSE was generally near 1 mg/L and *B* was relatively close to zero. No model dominated across all seven forecast accuracy measures. However, the performance of the existing ANN0 was exceeded in each forecast measure by at least one model tested in this study.

Table 3 | All models: training forecast parameters. Best performance is in bold

Model	RMSE	NASH	B	FAR	CSI	HSS	CB
ANN0	0.920	0.670	0.006	0.444	0.313	0.409	0.750
ANN1	0.908	0.701	**0.000**	0.500	0.368	0.461	**1.167**
ANN2	0.787	0.787	−0.002	0.437	0.474	0.579	1.333
ANN3	0.726	0.827	−0.002	0.429	0.444	0.550	**1.167**
ANN4	**0.699**	**0.840**	−0.002	0.429	0.444	0.550	**1.167**
ANN5	0.710	0.834	−0.002	0.400	**0.500**	**0.609**	1.250
EPR1	1.092	0.679	−0.048	0.500	0.200	0.268	0.500
EPR2	0.991	0.736	−0.012	**0.375**	0.333	0.440	0.667
NBM1	N/A	N/A	N/A	0.400	0.214	0.297	0.417
NBM2	N/A	N/A	N/A	0.600	0.286	0.347	1.250

Table 4 | All models: testing forecast parameters. Best performance is in bold

Model	RMSE	NASH	B	FAR	CSI	HSS	CB
ANN0	1.022	0.577	−0.020	0.333	0.333	0.455	0.600
ANN1	1.049	0.576	−0.027	0.250	**0.500**	**0.630**	0.800
ANN2	**0.935**	0.705	0.137	0.250	**0.500**	**0.630**	0.800
ANN3	0.939	0.704	0.119	0.333	0.333	0.455	0.600
ANN4	0.969	0.672	0.030	0.250	**0.500**	**0.630**	0.800
ANN5	1.078	0.633	0.120	0.333	0.333	0.455	0.600
EPR1	1.170	0.659	−0.166	**0.200**	0.364	0.496	0.500
EPR2	1.010	**0.742**	**− 0.017**	0.333	0.333	0.455	0.600
NBM1	N/A	N/A	N/A	0.333	0.200	0.300	0.333
NBM2	N/A	N/A	N/A	0.700	0.188	0.234	**1.111**

Additionally, the data mining models applied in this study could be used in determining the most relevant inputs and the best-fit shape of the prediction Equation (EPR), or in uncertainty analysis and ranking input variables (NBM).

CONCLUSIONS

Artificial neural networks (ANN), evolutionary polynomial regression (EPR) and naive Bayes model (NBM) were applied to predict weekly fluctuations of nitrate-N concentration at an agricultural watershed in central Illinois. Those predictions are critical in daily operations of the water supply utilities in the region. The models were compared using seven performance evaluation criteria. While all the models in this study produced smaller standard error compared with the previous studies, the results also demonstrated that none of the models was superior in all seven criteria, suggesting a multi-tool approach.

Nitrate-N prediction accuracy potentially could be increased by using hydro-meteorological forecasts, spatially distributed model inputs or by separating surface and base flows. In such relationships with increasing complexity, data mining tools, such as those presented in this study, could yield more accurate and precise forecasts. These tools also could be used in determining the relevant inputs, type of relationship and model size, and to assist water managers in selecting monitoring sites and variables, as well as determining observation frequency for nitrate-N and other water quality parameters.

ACKNOWLEDGEMENTS

This research was partially supported by the Faculty Fellowship Program, jointly funded by the National Center for Supercomputer Applications (NCSA) and the University of Illinois at Urbana-Champaign, Illinois. Radha Nandkumar, the Faculty Fellowship Program Director, served as a programmatic and technical liaison with NCSA. Also, this research was partially supported by the Illinois Indiana Sea Grant. We would like to thank Dr. Phil Mankin and Dr. Brian Miller for their continuous support. The data used in this study were collected under a monitoring project supported by the City of Decatur. The authors would also like to acknowledge the contribution of Laura Keefer, Hydrologist at the Illinois Water Survey for providing data, and Lisa Sheppard for editing.

REFERENCES

Amenu, G. G., Markus, M., Kumar, P. & Demissie, M. 2007 Hydrologic applications of minimal resource allocation network (MRAN) algorithm. *J. Hydrol. Eng.* **12** (1), 124–129.

Bajcsy, P., Han, J., Liu, L. & Young, J. 2004 Survey of bio-data analysis from data mining perspective. In *Data Mining Bioinformatics* (ed. J. Wang, M. Zaki, H. Toivonen & D. Shasha), pp. 9–39. Springer-Verlag, Berlin.

Bajcsy, P., Kooper, R., Marini, L., Clutter, D. & Markus, M. 2006 Visualization and data mining tools applied to algal biomass prediction in Illinois streams. In *Hydroinformatics 2006–Proceedings of the 7th International Conference on Hydroinformatics, Nice, France* (ed. P. Gourbesville, J. Cunge, V. Guinot & S.-Y. Liong), pp. 926–933. Research Publishing Services, Chennai, India.

Benedetti, A., Lopez, P., Moreau, E., Bauer, P. & Venugopal, V. 2005 Verification of TMI-adjusted rainfall analyses of tropical cyclones at ECMWF using TRMM precipitation radar. *J. Appl. Meteorol.* **44** (11), 1677–1690.

Bobbin, J. & Recknagel, F. 2001 Inducing explanatory rules for the prediction of algal blooms by genetic algorithms. *Environ. Int.* **27** (2–3), 237–242.

Bowden, G. J., Dandy, G. C. & Maier, H. R. 2005*a* Input determination for neural network models in water resources applications. part 1. background and methodology. *J. Hydrol.* **301** (1–4), 75–92.

Bowden, G. J., Maier, H. R. & Dandy, G. C. 2005*b* Input determination for neural network models in water resources applications. part 2. case study: forecasting salinity in a river. *J. Hydrol.* **301** (1–4), 93–107.

Bowden, G. J., Nixon, J. B., Dandy, G. C., Maier, H. R. & Holmes, M. 2006 Forecasting chlorine residuals in a water distribution system using a general regression neural network. *Math. Comput. Model.* **44** (5–6), 469–484.

Cohn, T. A., Caulder, D. L., Gilroy, E. J., Zynjuk, L. D. & Summers, R. M. 1992 The validity of a simple statistical model for estimating fluvial constituent loads: an empirical study involving nutrient loads entering Chesapeake Bay. *Water Resour. Res.* **28** (9), 2353–2363.

Davidson, J. W., Savic, D. A. & Walters, G. A. 1999 Method for the identification of explicit polynomial formulae for the friction in turbulent pipe flow. *J. Hydroinf.* **1** (2), 115–126.

Davidson, J. W., Savic, D. A. & Walters, G. A. 2000 Approximators for the Colebrook–White formula obtained through a hybrid regression method: computational methods in water resources. In *Computational Methods, Surface Water Systems and Hydrology* (ed. L. R. Bentley, J. F. Sykes, C. A. Brebbia, W. G. Gray & G. F. Pinder), Vol. 2, pp. 983–989. Balkema, Rotterdam.

Doglioni, A., Giustolisi, O., Savic, D. A. & Webb, B. W. 2008 An investigation on stream temperature analysis based on evolutionary computing. *Hydrol. Process.* **22** (3), 315–326.

Eder, B., Kang, D., Mathur, R., Yu, S. & Schere, K. 2006 An operational evaluation of the Eta-CMAQ air quality forecast model. *Atmos. Environ.* **40**, 4894–4905.

French, M. N., Krajewski, W. F. & Cuykendall, R. R. 1992 Rainfall forecasting in space and time using a neural network. *J. Hydrol.* **137** (1–4), 1–31.

Giustolisi, O. & Laucelli, D. 2005 Improving generalization of artificial neural networks in rainfall–runoff modelling. *Hydrol. Sci. J.* **50** (3), 439–457.

Giustolisi, O. & Savic, D. A. 2006 A symbolic data-driven technique based on evolutionary polynomial regression. *J. Hydroinf.* **8** (3), 207–222.

Giustolisi, O. & Simeone, V. 2006 Optimal design of artificial neural networks by a multi-objective strategy: groundwater level predictions. *Hydrol. Sci. J.* **51** (3), 502–523.

Giustolisi, O., Doglioni, A., Savic, D. A. & Webb, B. 2007 A multi-model approach to analysis of environmental phenomena (special issue). *Complex. Integr. Resour. Manage. Environ. Modell. Softw.* **5** (22), 674–682.

Giustolisi, O., Doglioni, A., Savic, D. A. & di Pierro, F. 2008 An evolutionary multi-objective strategy for the effective management of groundwater resources. *Water Resour. Res.* **44**, W01403 (doi:10.1029/2006WR005359).

Goldberg, D. E. 1989 *Genetic Algorithms in Search, Optimization and Machine Learning*, 1st edition. Addison-Wesley Longman, New York.

Guo, Y., Markus, M. & Demissie, M. 2002 Uncertainty of nitrate-N load computations for agricultural watersheds. *Water Resour. Res.* **38** (10), 3-1–3-12.

Haklander, A. J. & Van Delden, A. 2003 Thunderstorm predictors and their forecast skill for The Netherlands. *Atmos. Res.* **67** (8), 273–299.

Hall, M. J. & Minns, A. W. 1999 Classification of hydrologically homogeneous regions. *Hydrol. Sci. J.* **44** (5), 693–704.

Jain, S. K. 2008 Development of integrated discharge and sediment rating relation using a compound neural network. *J. Hydrol. Eng.* **13** (3), 124–131.

Keefer, L. & Demissie, M. 2000 *Watershed Monitoring for the Lake Decatur Watershed 1998–1999*. Illinois State Water Survey Contract Report 2000–2006, Champaign, IL.

Kohonen, T. 1988 An introduction to neural computing. *Neural Netw.* **1**, 3–16.

Koza, J. R. 1992 *Genetic Programming: On the Programming of Computers by Natural Selection*. MIT Press, Cambridge, MA.

Kumar, P., Alameda, J., Bajcsy, P., Folk, M. & Markus, M. 2006 *Hydroinformatics: Data Integrative Approaches in Computation, Analysis, and Modeling*. CRC Press, Boca Raton, FL.

Maier, H. R. & Dandy, G. C. 1996 The use of artificial neural networks for the prediction of water quality parameters. *Water Resour. Res.* **32**, 1013–1022.

Markus, M. 2005 Issues in designing automated minimal resource allocation neural networks. In *IEEE Proceedings of the International Joint Conference on Neural Networks*, **5**, pp. 2671–2673.

Markus, M. 2006 Artificial neural networks. In *Hydroinformatics: Data Integrative Approaches in Computation, Analysis, and Modeling* (ed. P. Kumar, J. Alameda, P. Bajcsy, M. Folk & M. Markus), pp. 411–438. CRC Press, Boca Raton, FL.

Markus, M., Tsai, C. W.-S. & Demissie, M. 2003 Uncertainty of weekly nitrate-nitrogen forecasts using artificial neural networks. *J. Environ. Eng.* **129** (3), 267–274.

Mathworks, Inc. 2007 *MATLAB–The Language of Technical Computing*. Available at: http://www.mathworks.com/products/matlab (August 14, 2008).

Mishra, A., Ray, C. & Kolpin, D. W. 2004 Use of qualitative and quantitative information in neural networks for assessing agricultural chemical contamination of domestic wells. *J. Hydrol. Eng.* **9** (6), 502–511.

Moradkhani, H., Hsu, K.-L., Gupta, H. V. & Sorooshian, S. 2004 Improved streamflow forecasting using self-organizing radial basis function artificial neural networks. *J. Hydrol.* **295** (1–4), 246–262.

Muttil, N. & Lee, J. H. W. 2005 Genetic programming for analysis and real-time prediction of coastal algal blooms. *Ecol. Modell.* **189** (3–4), 363–376.

Ochoa-Rivera, J. C., García-Bartual, R. & Andreu, J. 2002 Multivariate synthetic streamflow generation using a hybrid model based on artificial neural networks. *Hydrol. Earth Syst. Sci.* **6** (4), 641–654.

Ochoa-Rivera, J. C., Andreu, J. & García-Bartual, R. 2007 Influence of inflows modeling on management simulation of water resources system. *J. Water Resour. Plann. Manage.* **133** (2), 106–116.

Peng, F., Schuurmans, D. & Wang, S. 2004 Augmenting naive Bayes classifiers with statistical language models. *Inf. Retr.* **7** (3–4), 317–345.

Roebber, P. J., Bruening, S. L., Schultz, D. M. & Cortinas, J. V., Jr. 2002 Improving snowfall forecasting by diagnosing snow density. *Weather Forecast.* **18** (2), 264–287.

Salas, J. D., Markus, M. & Tokar, A. S. 2000 Streamflow forecasting based on artificial neural networks. In *Artificial Neural Networks in Hydrology* (ed. R. S. Govindaraju & A. R. Rao), pp. 23–51. Kluwer, Amsterdam.

Sharma, V., Negi, S., Rudra, R. & Yang, S. 2003 Neural networks for predicting nitrate-nitrogen in drainage water. *Agric. Water Manage.* **63** (3), 169–183.

Stenemo, F., Lindahl, A. M. L., Gardenas, A. & Jarvis, N. 2007 Meta-modeling of the pesticide fate model MACRO for groundwater exposure assessments using artificial neural networks. *J. Contam. Hydrol.* **93** (1–4), 270–283.

Suen, J. -P. & Eheart, J. W. 2003 Evaluation of neural networks for modeling nitrate concentrations in rivers. *J. Water Resour. Plann. Manage.* **129** (6), 505–510.

Thandaveswara, B. S. & Sajikumar, N. 2000 Classification of river basins using artificial neural network. *J. Hydrol. Eng.* **5** (3), 290–298.

Tokar, S. & Markus, M. 2000 Precipitation-runoff modeling using artificial neural networks and conceptual models. *J. Hydrol. Eng.* **5** (2), 156–161.

USEPA (United States Environmental Protection Agency) 1991 National Primary Drinking Water Regulations: Final Rule. 40 CFR Parts 141, 142 and 143. *Fed. Regist.* **56**(20), 3526–3597.

Yu, C., Northcott, W. J. & McIsaac, G. F. 2004 Development of an artificial neural network for hydrologic and water quality modeling of agricultural watersheds. *Trans. ASAE* **47** (1), 285–290.

Zhang, B. & Govindaraju, R. S. 2003 Geomorphology-based artificial neural networks (GANNs) for estimation of direct runoff over watersheds. *J. Hydrol.* **273** (1–4), 18–34.

First received 29 August 2008; accepted in revised form 12 June 2009. Available online 11 January 2010