November 2007

Source tracking of microbial intrusion in water systems using artificial neural networks

Minyoung Kim
University of Nebraska-Lincoln, mkim4@unl.edu

Christopher Y. Choi
University of Arizona, cchoi@arizona.edu

Charles P. Garba
University of Arizona

WATER RESEARCH ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Available at www.sciencedirect.com

ScienceDirect

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

Source tracking of microbial intrusion in water systems using artificial neural networks

Minyoung Kim[a,*], *Christopher Y. Choi*[b], *Charles P. Gerba*[c]

[a]*Agroecosystem Management Research Unit, USDA-ARS, 120 Keim Hall, East Campus, University of Nebraska, Lincoln, NE 68583, USA*
[b]*Department of Agricultural and Biosystems Engineering, University of Arizona, Tucson, AZ 85721, USA*
[c]*Department of Soil, Water, and Environmental Science, University of Arizona, Tucson, AZ 85721, USA*

ARTICLE INFO

Article history:
Received 9 January 2007
Received in revised form
22 September 2007
Accepted 27 September 2007

Keywords:
Source identification
Microbial intrusion
Artificial neural networks
Backpropagation
Generalized regression neural network

ABSTRACT

A "what-if" scenario where biological agents are accidentally or deliberately introduced into a water system was generated, and artificial neural network (ANN) models were applied to identify the pathogenic release location to isolate the contaminated area and minimize its hazards. The spatiotemporal distribution of *Escherichia coli* 15597 along the water system was employed to locate pollutants by inversely interpreting transport patterns of *E. coli* using ANNs. Results showed that dispersion patterns of *E. coli* were positively correlated to pH, turbidity, and conductivity ($R^2 = 0.90$–0.96), and the ANN models successfully identified the source location of *E. coli* introduced into a given system with 75% accuracy based on the pre-programmed relationships between *E. coli* transport patterns and release locations. The findings in this study will enable us to assess the vulnerability of essential water systems, establish the early warning system and protect humans and the environment.

1. Introduction

Sources for biological agents causing health-related risks range from naturally occurring contaminants to those accidentally or intentionally dispersed by humans. Many potential entry points exist, including waste disposal processes (toilet flushing), cleaning (e.g., bathing, hand washing), and indoor/outdoor sanitary activities (e.g., kitchen sink use, storm water collection, irrigation with reclaimed water), and water distribution systems (e.g., fire hydrants, storage tanks, irrigation channels) (Decker, 1990; Mark et al., 1998; Choi et al., 2003). Since the terrorist attacks of September 11, 2001, large and small utilities have also been concerned about the deliberate contamination of water distribution systems with chemical, radiological, and biological agents. When real-time sensors are available for water systems, the use of early warning algorithms and prediction models based on networked real-time data are imperative.

The intrusion of biological agents into water systems can pose serious public health risks because these agents cannot be easily detected and can remain hidden until a widespread contamination exists. Most cases with known causes have demonstrated considerable delay by authorities to perceive a risk and respond to the situation. For instance, the largest waterborne outbreak in the US resulted in massive illness among 403,000 people in Milwaukee (Ford and MacKenzie, 2000). Ford and MacKenzie (2000) noted that water was contaminated for at least 2 weeks before oocysts were identified in stool samples and water contamination was finally suspected. In instances like this, it is difficult to manage the contamination event rapidly and efficiently as well as to minimize further health impacts without first knowing the source and location of pathogen release. If the geological location and release records of biological contaminants were known at a site, further transmission of

*Corresponding author. Tel.: +1 402 472 9298; fax: 1 402 472 0516.
E-mail address: Minyoung.Kim@ars.usda.gov (M. Kim).
0043-1354/$ - see front matter © 2007 Published by Elsevier Ltd.
doi:10.1016/j.watres.2007.09.032

contaminants could be intercepted with minimal health risk. This early detection results in faster completion of remediation methods and risk assessments for the contamination site.

Identification of contaminant sources is not an easy task. At present, many cost-effective sensors for real-time monitoring of water quality exist that allow for a chemical fingerprint. These sensors measure important water quality parameters such as pH, free chlorine, total organic carbon, and total oxygen. The fingerprints based on these parameters are designed to alert utilities of a possible intrusion event in real-time when a contaminant is injected into the distribution system. Additionally, a few technologies are commercially available and are being evaluated for the identification of specific biological agents for water. The use of monitoring and prediction tools in conjunction with these sensors will greatly assist in decision-making within a short time period in the event of water contamination.

Artificial neural networks (ANNs) were proposed in this study as a decision-making tool, along with a computational flow model (Model Of Urban Sewers, or MOUSE) supported by experimental data. The literature on the applications of ANNs is extensive and excellent summaries in science and engineering can be found in Basheer and Hajmeer (2000) and Maier and Dandy (2000).

A key idea for the present study is that unique dispersion patterns extracted from a contaminant's release curve provide valuable information to track the time and distance of the release. For example, the plume of a contaminant introduced at a distance far from the monitoring site would have a longer travel time, greater spread of contaminant plume, and a lower peak concentration than a contaminant introduced closer to the monitoring site. Interpreting patterns is a critical factor for the development of a reliable model that can be used to isolate the intrusion location of contaminants in a water system.

The specific objectives of this study are (1) estimation of time series of *E. coli* concentration data using water quality parameters (pH, turbidity, and conductivity); (2) pattern extraction from *E. coli* breakthrough curves; (3) training ANNs to recognize the uniqueness of each microbial transport pattern depending on its injection location; (4) comparison of pre-recognized patterns with unidentified transport patterns; and (5) determination of contaminant release locations.

2. Materials and methods

2.1. Description of field study and hydraulic monitoring

A scale model of an open water system was designed and built at the Agricultural Research Center at the University of Arizona in Tucson, AZ (Fig. 1). The key experimental design parameters (e.g., dimensions, flow speed, etc.) were carefully examined with MOUSE® (DHI Inc., Portland, OR, USA) prior to the model's construction. The main, sub-main, and water-feeding pipes were constructed from interconnected circular PVC pipes of 10.16 and 1.27 cm diameter.

Solenoid valves regulated the water flow for each inlet point. The valves were controlled using a CR10X datalogger (Campbell Scientific Incorporation, Logan, UT). The datalogger was programmed using LoggerNet (Campbell Scientific Incorporation, Logan, UT) and used both as a data-storage device for pressure transducer readings and also as a controller for the solenoid valves. Steady-state flow conditions ($q = 31.56$, 63.10, 94.65, and 126.20 cm^3/s) were generated by solenoid valves and confirmed by a calibrated pressure transducer at the outlet point.

2.2. Transport phenomena of microorganisms

Prior to the microbial transport study using *E. coli* 15597, a conservative tracer test using sodium chloride (NaCl) was conducted to characterize the hydraulic and transport properties of the system. The fecal coliform indicator bacterium, *E. coli* 15597, was used because it is associated with enteric pathogens. A background test was conducted prior to the release of the *E. coli* solution in order to confirm the absence of any naturally occurring microorganisms.

The microbial transport study procedure is as follows: 500 mL of *E. coli* suspension ($\sim 10^8$ colony-forming units per milliliter, cfu/mL) was introduced into inlet 1 (In 1) for about 40 s. At specific time intervals ($t = 120$, 140, 160, 210, 240, 360, 600, 1800, 3600, and 7200 s), 30 mL samples of water from the outlet site were collected in pre-sterilized 50 mL plastic centrifuge tubes. This procedure was then repeated for the remaining inlet points (In 2, 3, 4, and 5). All samples were stored in an ice chest and immediately delivered to the Environmental Microbiology Laboratory at the University of Arizona for processing. All samples containing *E. coli* were diluted and cultured onto plates of eosin methylene blue (EMB) agar using the spread plate method (American Public Health Association, 1991) under aseptic conditions. A 0.1 mL aliquot of selected dilutions of the sample was uniformly spread on top of the EMB agar with a sterile glass rod. After plating, the samples were incubated overnight at 35 °C, allowing the bacteria to multiply into isolated colony-forming units (cfu) (Maier et al., 2000). In addition, pH and turbidity readings from each water sample were measured within 24 h, using a Corning 445 pH meter and a Hach 2100AN turbidimeter (San Diego, CA, USA), respectively. All measurements of microbial concentration, pH, turbidity, and EC were repeated three times and their average values were used to plot the transport patterns over time.

2.3. Prediction of microbial transport behavior using ANNs

ANNs estimate an output vector, Y, for a given input vector X, such as $Y = g(X)$ for a given function $g(\bullet)$ and then provide an approximation to any function of the input vector X (Gallant and White, 1988; Irie and Miyake, 1988). This function of ANNs expressed using backpropagation (BP) and generalized regression neural network (GRNN) was applied in this study to define the relationship between the microbial transport and its corresponding water quality parameters.

Because each input and output variable measured different phenomena, each had very different ranges, i.e., conductivity

Please cite this article as: Kim, M., et al., Source tracking of microbial intrusion in water systems using artificial neural networks. Water Res. (2007), doi:10.1016/j.watres.2007.09.032

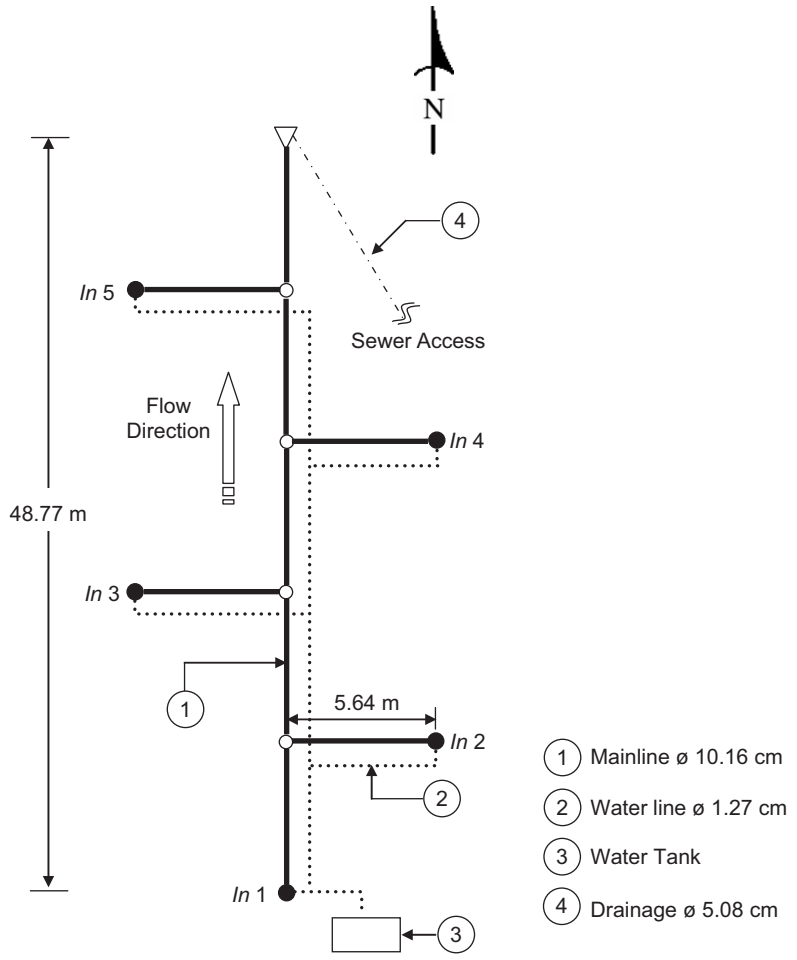

Fig. 1 – Plan view of the experimental layout using interconnected PVC pipes with various diameters (slope 0.7%).

(76–350), pH (7.095–8.350), turbidity (0.139–11.555), and *E. coli* concentrations (10^0–10^8). As a pre-processing step, all input (three water quality parameters) and output (*E. coli* concentrations) data require one normalization (scaling) within a uniform range (0–1). Without rescaling, the network becomes highly sensitive to the processing element (PE) with a larger value, because the magnitude was related to its unit of measurement, not the relative importance.

After normalizing values, selecting the most appropriate input–output datasets is another important consideration in determining the success or failure of a neural network application. Approximately 90% of the time and effort involved in computation was devoted to this step (Warner and Misra, 1996). Data to be used for training should be sufficiently large to cover the possible known variations in the problem domain (Swingler, 1996). Therefore, a total of 488 data were divided into three subsets, and 72% (351), 8% (39), and 20% (98) data were used for training, validating, and testing networks, respectively. All data were randomly divided into $k = 10$ subsets of equal size, and different network architecture and training parameters were selected and trained. Each time one of the subsets from the training was left out during the training, but the omitted subset was used for validation. This procedure was repeated until no further decrease in error occurred among 10 subsets. After training and validation, the network architecture having the smallest error over 10 subsets was selected and evaluated using the test set to complete the procedure.

There are numerous factors used to achieve the best model performance for BP, which include the number of hidden layers, the number of hidden PEs, transfer function (sigmoid, tan-sigmoid, etc.), learning algorithms (Delta, extended DBD, etc.), and learning parameters (learning rate, momentum factor, and initial weights). Detailed information about each parameter (definition, function, range, etc.) is provided in Basheer and Hajmeer (2000) and Maier and Dandy (2000).

Depending on the problem being solved, the success of training varies with the selected factors, and a trial-and-error procedure is normally preferred. The root mean of square error and the classification ratio (correlation coefficient, *R*), which assessed the fitness of each performance, were provided by NeuralWorks Professional II/PLUS (Carnegie, PA, USA) software, version 5.22.

Contrary to the BP algorithm, GRNN is capable of approximating any arbitrary function from historical data, as its name implies. The foundation of the GRNN operation is essentially based on the theory of nonlinear (kernel) regression. The smoothing factor, σ, is the most important computing parameter for the GRNN's performance. Theoretically, it is not possible to determine the true σ because the

Please cite this article as: Kim, M., et al., Source tracking of microbial intrusion in water systems using artificial neural networks. Water Res. (2007), doi:10.1016/j.watres.2007.09.032

underlying parent distribution is not known. However, the optimum σ for a GRNN built from a training data set can be automatically approximated using the "holdout method" (Specht, 1991; Chtioui et al., 1999). In this study, the σ value began at 0.005 with a certain interval, and the final σ value was determined by the trial-and-error method. GRNN was created with the Neural Network Toolbox for MATLAB® (The Mathworks, Natick, MA, USA).

2.4. *Source tracking with ANNs*

The simulation of the output state of a model based on known input variables is called "the forward problem." By contrast, the inverse problem is intended to find the unknown input variables that give rise to a partially known output state (Liu and William, 1999). In the source identification of a contaminant, the injection locations are unknown. In order to represent data more efficiently, the most important features in the data set must be extracted from microbial release curves. The crucial task is feature extraction, which uses characteristic features to find unknown regularities, meaningful categorizations, and patterns in the presented input data. In other words, transport characteristics obtained from a breakthrough curve when *E. coli* was injected into In 1 must be such that points in a single class (In 1) are close to each other and points from different classes are farther apart (characteristics from releasing event through In 2, 3, 4, and 5).

A total of 13 carefully inspected parameters were extracted, characterized, and categorized depending on the releasing location (In 1, 2, 3, 4, and 5). Parameters included the following: hydraulic factor (flow rate), geometric parameters (distance from inlet to outlet), and microbial quality data (total amount of *E. coli* introduced, time variable concentrations at time = 120, 140, 160, 210, 240, 360, 600, 1800, 3600, and 7200 s) under three different flow conditions ($q = 31.56$, 63.10, and 94.65 cm^3/s).

While many different types of neural networks have been developed, a BP design was chosen because of its demonstrated effectiveness at classifying nonlinear data sets.

3. Results and discussion

3.1. *Field data*

Each experiment was repeated twice, and all readings were averaged and normalized for a better description as shown in Fig. 2. An *E. coli* suspension was injected into In 1, which is the farthest inlet point from the outlet. Time-dependent water samples were used to quantify *E. coli* concentrations, pH, and turbidity levels over time. Shortly after the injection event, significant changes were observed from the baseline values of pH and turbidity depending on the flow conditions ($q = 31.56$, 63.10, 94.65, and 126.20 cm^3/s). Measurement errors were observed at 63.10 cm^3/s, and the *E. coli* concentration curve began about 15 s earlier. Turbidity and pH readings were measured within 24 h after sample collection. When using tryptic soy broth as a growth medium for *E. coli*, metabolic

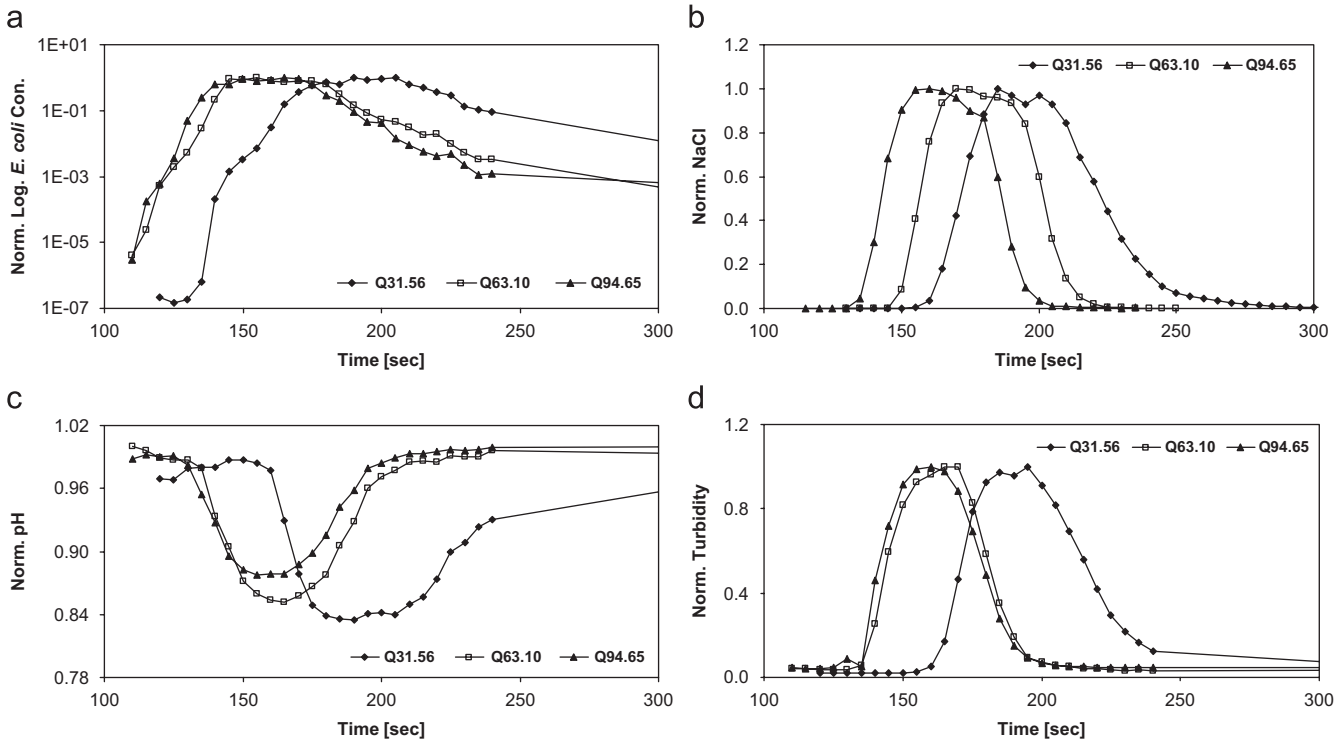

Fig. 2 – Plots of time-series *E. coli* and other water quality parameters under various flow conditions: (a) *E. coli*, (b) conductivity, (c) pH, and (d) turbidity.

Please cite this article as: Kim, M., et al., Source tracking of microbial intrusion in water systems using artificial neural networks. Water Res. (2007), doi:10.1016/j.watres.2007.09.032

acids produced by bacteria and microbial activities influenced the change in pH and turbidity to a certain degree over time.

3.2. *Model performance*

Two algorithms, BP and GRNN, were applied and compared to predict the transport behavior of *E. coli* using the three water quality parameters. Time-series measurements for *E. coli* concentration and water quality parameters (pH, turbidity, and conductivity) were entered into BP and GRNN as input and output variables, respectively. Simulation results are plotted in Fig. 3 using experimental measurements from the three different flow conditions.

Several data points were excluded from the curves because pH and turbidity readings were not sensitive enough to recognize changes in *E. coli* concentrations, creating unnecessary errors. Partial data sets significantly improved the model performance: $R^2 = 0.61–0.92$ (BP) and 0.60–0.90 (GRNN)

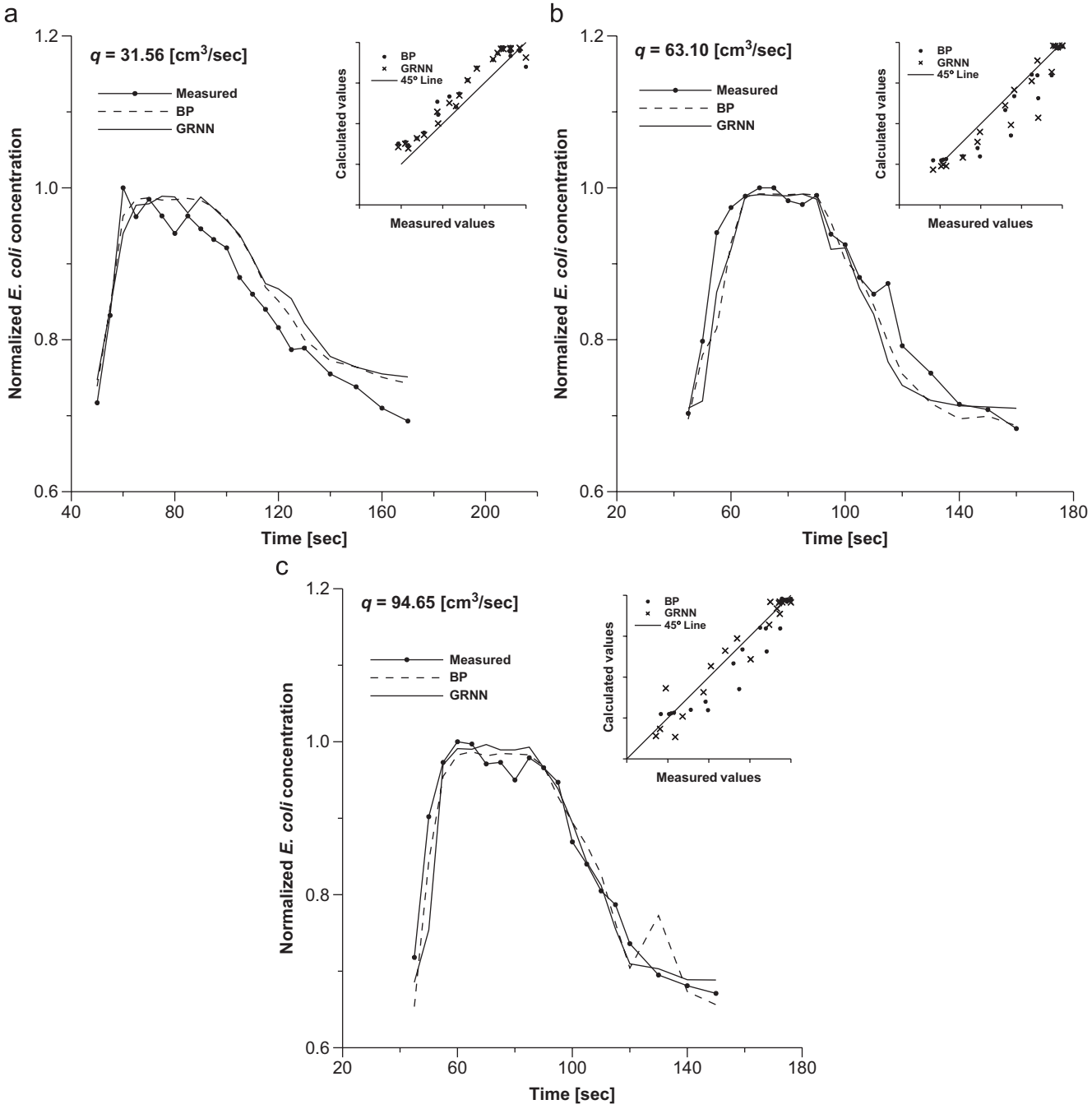

Fig. 3 – Prediction of *E. coli* concentrations with the aid of its surrogates, pH, turbidity, and conductivity data under various flow rate conditions: (a) 31.56, (b) 63.10, and (c) 94.65 cm^3/s.

Please cite this article as: Kim, M., et al., Source tracking of microbial intrusion in water systems using artificial neural networks. Water Res. (2007), doi:10.1016/j.watres.2007.09.032

for the entire data sets and $R^2 = 0.92–0.96$ (BP) and 0.90–0.92 (GRNN) for the partial data sets. As shown here, the overall model resulted in a predicted quantity of *E. coli* that consistently matched the actual concentrations fairly well. This indicates that conductivity, pH, and turbidity are positively correlated with *E. coli* concentrations.

The BP algorithm showed the best performance with a feed-forward connection type and 3–1–0–1 (numbers of PEs in the input layer, the first hidden layer, the second hidden layer, and the output layer, respectively) architecture. The learning rate and momentum were experimentally determined as 0.3 and 0.4, respectively. The optimal smoothing factor for GRNN was determined as 0.015. The performance of each algorithm is summarized in Table 1.

In the source identification problem, the challenge was the feature extraction of concentration profiles. Useful parts of each curve were travel time, time of peak concentration, amount of peak concentration, and variation of concentration over time. In contrast to the first and prominent peak of the curve used as the feature extraction, it was difficult to separate pattern features from the rest of the curve. For this experiment, unique transport patterns of *E. coli* injected from different inlets (In 1, 2, 3, 4, and 5) under three different flow conditions (31.56, 63.10, and 94.65 cm^3/s) were categorized based on the inlet location. A total of 15 variables were used for training and testing purposes to classify the transport characteristics. The best model performance was achieved using the BP algorithm with the architecture of 15–1–0–1 (numbers of PEs in each layer), the Delta learning rule, the TanH transfer function, and 50,000 iterations by the trial-and-error method. Tables 2 and 3 show the classification results from the best-performing architecture of BP with the aforementioned parameters. The classification results were reasonably good. In the training process, only two out of 11 variables failed to classify into the correct class (82% correct identification). When four input variables were used to test the network architecture, only one case failed (75% correct identification).

4. Conclusions

A scale model of an open water system was designed using a computer model (MOUSE®) and constructed to produce the necessary experimental data. Pre-defined concentrations of NaCl and *E. coli* were introduced into each inlet, and water samples were collected from the outlet at designated time intervals. Water quality parameters (pH, conductivity, and turbidity) were measured to assess their feasibility as relevant parameters of *E. coli* concentrations. Two algorithms in ANNs (BP and GRNN) were compared to evaluate their potential capabilities as universal approximating and decision-making tools.

A positive correlation between *E. coli* concentration and the three selected water quality parameters was demonstrated, indicating that ANNs successfully estimated *E. coli* concentrations as a function of pH, turbidity, and conductivity. For source identification of a contaminant, unique characteristics from *E. coli* concentration profiles were extracted and classified into each classifier (inlet location). The approach used in this study proved promising. Nevertheless, additional studies are needed to ascertain which combinations of parameters can be universally applied.

This study clearly proved that source locations and release histories at water system sites can be identified and reconstructed. This will allow for remediation system design and risk assessment studies to be completed within a shorter

Table 3 – Testing results for source identification problem

Classifier	Testing set				Accuracy
Target	2	3	4	5	75%
Computed	1	3	4	5	
Result[a]	IC	C	C	C	

[a] Correctly identified (C), incorrectly identified (IC).

Table 1 – Compared performance of BP and GRNN models against field data

	BP			GRNN		
q (cm^3/s)	31.56	63.10	94.65	31.56	63.10	94.65
Mean square error (MSE)	0.0012	0.0015	0.0010	0.0016	0.0016	0.0015
Coefficient of determination (R^2)	0.960	0.901	0.927	0.929	0.935	0.943

Table 2 – Training results for source identification problem

Classifier	Training set											Accuracy
Target inlet location	1	1	1	2	2	3	3	4	4	5	5	87%
Computed inlet location	2	1	1	1	2	3	3	4	4	5	5	
Result[a]	IC	C	C	IC	C	C	C	C	C	C	C	

[a] Correctly identified (C), incorrectly identified (IC).

Please cite this article as: Kim, M., et al., Source tracking of microbial intrusion in water systems using artificial neural networks. Water Res. (2007), doi:10.1016/j.watres.2007.09.032

period of time under various contamination events. This proposed methodology, which uses ANNs, has additional benefits that can be applied to any type of water distribution system (e.g., drinking water and agricultural irrigation systems) and water collection system (e.g., sewer systems). In addition, standard water quality models (e.g., MOUSE®) can be integrated with a geographic information system for accurate spatial analysis, enabling the model to better utilize the open database model along with hydraulic and water quality simulation results.

Resilient self-monitoring water infrastructure has become a new paradigm, and real-time sensors and event-monitors are being developed and tested by private companies, academic institutions, and local and federal agencies in the United States. These sensors and a sensor-network can provide specific threat information through appropriate linkages to a SCADA system. A decision-making computer model can help local utilities respond with adequate protective measures to mitigate any public health risk and isolate the contaminated areas for an immediate decontamination operation. If the threat is proven to be real, a key element in responding to the event is to identify the location and time of the contaminant release. Source-tracking algorithms based on ANNs can play a significant role as a part of the decision-making procedure.

REFERENCES

American Public Health Association, 1991. Standard Methods for Examination of Water & Wastewater. New York, NY.

Basheer, I.A., Hajmeer, M., 2000. Artificial neural networks: fundamentals, computing, design, and application. J. Microbiol. Methods 43, 3–31.

Choi, C.Y., Gerba, C.P., Riley, M., 2003. Environmental dispersion of biological agents in sewer systems. Final Report, DARPA Project no. 806345617.

Chtioui, Y., Panigraphi, S., Francl, L., 1999. A generalized regression neural network and its application for leaf wetness prediction to forecast plant disease. Chemometrics Intell. Lab. Syst. 48, 47–58.

Decker, R.G., 1990. Sewer line collapse at Prince and Oracle or how not to spend Labor Day weekend. Pima County Wastewater Management Department, PE.

Ford, T.E., MacKenzie, W.R., 2000. How safe is our drinking water? Despite technologic advances, waterborne disease is still a threat, Postgrad. Med. 108(4). Available from 〈http://www.postgradmed.com/issues/2000/09_00/editorial15sept.htm〉. Accessed June 2005.

Gallant, A.R., White, H., 1998. There Exists a Neural Network That Does Not Make Avoidable Mistakes. In: Proceeding of the International Conference on Neural Networks, vol. 1, pp. 657–666.

Irie, B., Miyake, B., 1988. Capabilities of three layer perceptrons. In: Proceeding of the IEEE Second International Conference on Neural Networks, vol. 1, pp. 641–648.

Liu, C., William, P.B., 1999. Application of inverse methods to contaminant source identification from aquitard diffusion profiles at Dover AFB, Delaware. Water Resour. Res. 35 (7), 1975–1985.

Maier, H.R., Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modeling issues and applications. Environ. Model. Software 15, 101–124.

Maier, R.M., Pepper, I.L., Gerba, C.P., 2000. Environmental Microbiology. Academic Press, San Diego, CA.

Mark, O., VanKalken, T., Rabbi, F., Albinsson, B., 1998. Risk analyses for sewer system based on numerical modeling and GIS. Saf. Sci. 30, 99–106.

Specht, D.F., 1991. A generalized regression neural network. IEEE Trans.Neural Networks 2 (6), 568–576.

Swingler, K., 1996. Applying Neural Networks: A Practical Guide. Academic Press, New York.

Warner, B., Misra, M., 1996. Understanding neural networks as statistical tools. Am. Stat. 50 (4), 284–293.

Please cite this article as: Kim, M., et al., Source tracking of microbial intrusion in water systems using artificial neural networks. Water Res. (2007), doi:10.1016/j.watres.2007.09.032