Neural network modelling of *Cryptosporidium* and *Giardia* concentrations in the Delaware River, USA

T.R. Neelakantan, G.M. Brion and S. Lingireddy
Dept of Civil Engineering, University of Kentucky, 161 Raymond Bldg., Lexington, Kentucky 40506-0281, USA (E-mail: gbrion@ engr.uky.edu)

**Abstract**  Artificial neural networks are brain-like structures used in mathematical modelling that excel in pattern recognition. In this research, a simple feed-forward artificial neural network, trained by error back-propagation algorithm, was used as a tool to relate peak *Cryptosporidium* and *Giardia* concentrations with other biological, chemical and physical parameters in surface water. Multiple water quality parameters at a water treatment plant intake on the Delaware River, New Jersey, USA, collected in 1996, were provided to the authors for recognition analysis. Water samples were classified as "background" and "above background" based on the concentrations of full and empty oocysts and cysts of *Cryptosporidium* and *Giardia*. The results of this preliminary effort were encouraging. Parameters significant to the identification of each protozoa were identified, eight for *Cryptosporidium* and seven for *Giardia* by a stepwise elimination technique. Data withheld from the model training was used to validate the trained models and evaluate the most effective internal architecture. In both cases, the best prediction performance was found when the number of internal nodes was twice that of the input parameters in single hidden-layer architecture.

Predictions for the classification of the verification data set resulted in no false-negatives (mis-prediction of above background protozoa concentrations) when the models were optimally trained.

**Keywords** Surface water; *Cryptosporidium*; *Giardia*; neural networks; water quality models

**Introduction**

Controlling surface water quality can effectively control the health risk associated with waterborne disease. However, complete elimination, or perfect control, of the multiple sources of pathogen inputs within a watershed is impossible. Therefore, on-line (continuous) monitoring systems are needed that identify when water quality is unacceptable or when peak pathogen contamination events are likely to occur. Since methods for the identification of pathogens are very costly and time consuming, surrogate quality measures must be used. Since no single surrogate measure has been shown to be consistently, significantly and statistically related to all pathogen presence in all surface waters, multiple parameter systems are being investigated for use in development of on-line alarms and eventual use by expert control systems. It has been suggested that in complex situations, dependent upon multiple, interrelated input variables, artificial neural networks may be appropriate tools for data analysis. This paper is an attempt in this direction, an exploration of neural networks to uncover which parameters are predictive of peak contamination events for encysted protozoa, *Cryptosporidium* and *Giardia*.

Low concentration doses of some pathogenic microorganisms can have acute effects on consumers, particularly the immunosuppressed. Water disinfecting processes are generally sufficient to remove the majority of microbial contaminants but some microorganisms survive disinfection, enter water distribution systems intact and are capable of infecting consumers. Encysted forms of *Cryptosporidium* and *Giardia* are among these hardy microorganisms. In the US, protozoa are the most commonly identified pathogen in waterborne disease outbreaks and cause the majority of hospitalisations (Craun *et al.*., 1998). Encysted protozoa have low infectious doses and cause illness ranging from mild gastroenteritis to prolonged, profuse and potentially life-threatening diarrhoea. Most of the time the
illness caused is self-limiting but for the immunocompromised infections can become chronic and life threatening. Although *Giardia* is the most commonly identified pathogen in waterborne outbreaks, *Cryptosporidium* is perceived to pose the greatest threat due to the lack of effective drug therapy. *Cryptosporidium* also gains the limelight due to media coverage of large outbreaks such as impacted Milwaukee in 1993. This was the largest outbreak of waterborne disease in the history of US causing about 400,000 cases of disease. The source of contamination was never clearly identified although animal husbandry as a non-point source was implicated (MacKenzie *et al.*, 1994). *Cryptosporidium* outbreaks have been reported in many other countries as well.

Both protozoa are ubiquitous in surface waters but the ultimate source of encysted protozoa is faeces from infected hosts. Contamination of a water supply can occur in two ways through the activity of animals in the watershed area of the water supply or by the introduction of sewage into the water supply. Studies have shown that, unlike many other pathogens, *Giardia* and *Cryptosporidium* are not host specific. This means that *Giardia* cysts excreted by animals can infect and cause illness in humans. Wild and domestic animals have been shown to be major contributors in contaminating water supplies. Additionally, several major outbreaks of waterborne *Giardia* have been found to be caused by sewage-contaminated water supplies. However, it is well known that protozoan concentrations increase with the flushing action of rainfall and peak during times of high runoff.

Testing for small numbers of encysted protozoa in surface waters has proven difficult. Analytical techniques are expensive, time-consuming, complicated and not always accurate. Therefore, defining surrogate indicators for protozoa is an active area of research. To date, neither organism has shown consistent correlation with other parameters (Jakubowski *et al.*, 1996) and no single surrogate has been adopted although several look promising. This is understandable when looking at the multiple sources of waterborne protozoa coupled with their longevity in the environment. Yet, protozoa have shown weak correlations with multiple, candidate surrogate water quality measures (LeChevallier *et al.*, 1991; Atherholt *et al.*, 1998). However, attempts to create predictive models have not succeeded and model performance suffers from watershed individuality, parameter inter-relatedness and analytical variability. In this research, an attempt was made to build a neural network predictive model that would use multiple parameters to create recognisable water quality “fingerprints” that would indicate if waterborne protozoa were present in numbers greater than normal. Neural networks were selected as a modelling tool because of their capability to capture non-linear relationships present in the data as well as the ability to self-train.

Feed forward, error back-propagation, neural networks work based on the principle of iterative curve fitting technique. These networks comprise a set of highly interconnected but simple processing units referred to as nodes. Neural networks learn through reviewing examples with known outcomes. As in curve fitting techniques, known data is used to train and validate the model before it is used for predictions. A database collected and used by Atherholt *et al.* (1998) in 1996 on the Delaware River at Trenton (40.2° N, 74.75° W) in New Jersey, USA was used for this study. The watershed that feeds this site extended over 17,527 km² and contained numerous human and animal sources of faecal pollution. Water sampling was done at the Trenton Water Works, New Jersey as previously described (Atherholt *et al.*, 1998). The database and all of the methods used to analyse and assemble values for encysted protozoa and other water quality parameters have been described in detail by Atherholt *et al.* (1998). A total of 72 d of observations were collected. Each observation in the database consisted of carefully measured physical, chemical and microbial parameters. Along with these commonly measured parameters, results for encysted protozoa were included. The method used to collect and detect encysted protozoa from the river was the EPA Information Collection Rule (ICR) filter cartridge method and, although this
method has been criticised for lack of precision, reliability, and high numbers of non-detects (Clancy et al. 1999), by consistently analysing 5 L equivalent volumes for every sample the error associated with the method was minimised.

**Neural networks**

The Artificial Neural Network (ANN) approach is a branch of artificial intelligence. The ANN is based on a model of the human neurological system that consists of basic computing elements (called neurons) interconnected together (Figure 1). With this parallel-distributed processing architecture, ANNs have proven to be very powerful computational tools that excel in pattern recognition and function approximation. As shown by Hornik et al. (1989), an ANN with sufficient complexity is capable of approximating any smooth function to any desired degree of accuracy. In addition, ANNs are computationally robust, having the ability to learn and to generalize from examples to produce meaningful solutions to problems even when the input data contain errors or are incomplete. If carelessly used, an ANN can easily “learn” irrelevant information (noises) in the system with the resulted ANN model being able to predict past incidents but unable to predict future events.

The ANN modelling framework has been used increasingly in various aspects of science and engineering because of its ability to model both linear and non-linear systems without the need to make any assumptions as are implicit in most traditional statistical approaches. Some of the problems in which ANNs have been used for include rainfall-runoff modelling, scheduling of hydroelectric power systems and river flow prediction. The ANN methodology has been applied also to forecast rainfall. Recently, neural networks were used for identifying non-point sources of microbial contamination (Brion and Lingireddy, 1999). A simple ANN of N input nodes, L hidden nodes and one output node (architecture of N-L-1) is shown in Figure 1 to illustrate this basic structure and terminology showing that an ANN consists of layers of artificial neurons or nodes. The nodes in a layer do not interconnect with other nodes in the same layer. In other words, connections occur only between layers and not within a layer. Three generic types of layers are presented in an ANN. The first layer, which receives input information, is called an input layer. The last layer, which produces output information, is called an output layer. Between the input and output layers are the hidden layers. Within an ANN, there can be one or more hidden layers. Information is transmitted through the connections between nodes in different layers. In a simple situation, the information, or signal, is passed forward only (Figure 1). This type of network is called a feed-forward network, or multi-layer feed-forward network.

![Figure 1](https://iwaponline.com/wst/article-pdf/43/12/125/428989/125.pdf)

**Figure 1** Three-layer error back-propagation neural network architecture
Mathematically, a three-layer neural network with $I$ input nodes, $J$ hidden nodes and $K$ output nodes, can be expressed as:

$$O_{P_k} = f_1\left(\sum_{j=1}^{J} w_{jk}^o f_2\left(\sum_{i=1}^{I} w_{ij}^h x_{P_i}\right)\right), \quad \forall k \in 1, 2, \ldots, K$$

(1)

where $O_{P_k}$ is the output from the $k$th node of the output layer of the network for the $P$th vector (data point); $x_{P_i}$ the inputs to the network for $P$th vector (data point); $w_{ij}^h$ the connection weight between $j$th node of the hidden layer and $k$th node of the output layer (see Figure 1); $w_{ij}^h$ the connection weight between $i$th node of the input layer and $j$th node of the hidden layers; and $f_1$ and $f_2$ are activation functions. The most commonly used activation function is a logistic sigmoidal function which has a form given below:

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

(2)

The main control parameters of a feed forward network are the connection weights. The process of estimating these parameters is known as training where optimal connection weights are determined by minimising an error function. The training algorithm by which the error at output layer is propagated to the previous layers thereby adjusting the weights in iterative process is called the error back-propagation algorithm. This back-propagation algorithm was used in this study. More details can be referred elsewhere (Masters, 1993).

The flexibility of an ANN comes from the fact that the number of hidden nodes, or even the number of hidden layers, can be easily changed to adjust the computational power of an ANN. This flexibility allows the modelling of complex systems even though there is little knowledge about the form of relationship between the independent and dependent variables. This, together with other advantages of ANNs, encourages the use of ANNs for this application. The greatest difficulty was determining the appropriate model inputs for such a complex problem. Though ANNs belong to the class of data-driven approaches, it is important to determine the dominant model inputs as this reduces the size of the network. Consequently there is a reduction of training times and increase of the generalisation ability of the network for a given data set. Sensitivity based judgement was used to reduce the number of model inputs in this case study.

**Methods**

In this study, *Cryptosporidium* and *Giardia* concentrations were modelled separately. Easily measurable water quality parameters were used as input variables for the model and the peak concentrations (see methods for definition of peak) of *Cryptosporidium* or *Giardia* were used as output. All the input parameter values to the neural network were rescaled to follow a maximum of 1 and minimum 0 by division with the maximum value. Rescaling, normalising and standardising are terms that are used more or less interchangeably depending on the customs within various fields. Standardising input variables has different effects on different training algorithms but is an important consideration for gradient descent methods such as back-propagation learning. Since not all the 72 d of observations in the database were complete for all parameters, 68 complete observations points were selected and used in the analysis. Using random numbers, the data points were mixed well to disturb their time sequence then 51 points were used for training and 17 points were used for validation of the trained model.

**Encoding peak and non-peak protozoa concentrations**

In this study, the values for oocyst concentrations were excerpted out of the database and reviewed in isolation to determine what cluster of concentrations represented background
or normal, levels of encysted protozoa at the sample site. Both numbers of oo(cysts) and condition of oo(cysts) (full or empty) were used to create the data classification scheme. For *Cryptosporidium* much of the analyses showed non-detect results or empty oocysts. Descriptive statistics revealed that the median, 25% and 75% percentiles for total numbers of oocysts were 0, 0 and 1 oocyst respectively. Using this information, observations were sorted into either background (0 or 1 empty oocyst) or above background (>1 oocyst with at least one oocyst having internal contents) concentrations. For example, an observation with 1 oocyst recovered from a 5 L equivalent volume, with one of those oocysts noted as having internal contents, would be output encoded as 1. Whereas, an observation of 1 oocyst without contents would be output encoded as 0.

For *Giardia*, which was present in higher numbers than *Cryptosporidium*, the data was split into background and above background levels in much the same way. The median number of total cysts (irrespective of contents) was 2 with the 75th percentile equal to 4 cysts. When analysing just the numbers of full cysts, the 75th percentile was equal to 1 full cyst. So it was decided that above background observations (encoded as 1) would be those days when there were 4 or more cysts detected with at least 1 cyst full. All other observations with less than 4 cysts were considered background (encoded as 0). There were only three observations that did not neatly fall out with this classification scheme. One observation had 2 cysts isolated and both contained contents. This was encoded as above background as it was reasoned that because analytical error can vary by as much as 100%, there could have easily been 4 cysts actually present. Two other observations had 3 cysts recovered with one showing contents. These were also encoded as 1 using the same rationale. Thus the encode 1 represented a peak concentration (75th percentile and above with possibly viable contents observed) and encode 0 represented non-peak or background concentration. This binary information was used as output of the model for training.

**Model training**

Trials were made with a forward feed, back-propagation trained, neural network with different combinations of water quality parameters as inputs. The following water quality parameters were evaluated as inputs to the neural network models: alkalinity, pH, turbidity, suspended solids, dissolved solids, river flow, precipitation, total coliphage, male-specific coliphage, total coliform, faecal coliform, *E. coli* and *Clostridium perfringens*. For each new combination of parameters, the training dataset was used to train the model until the mean squared error was significantly small, i.e. in the order of 0.05. After training, the input parameters for the validation set were fed into the trained model and predictions of the oocyst classification category, or output values, produced. Predictions in the range 0–0.3 were accepted as background (0). Predictions in the range 0.7–1.0 were accepted as above background (1). Predictions in the range between 0.3 and 0.7 were not accepted as either background or above background but automatically classified as a mis-prediction. Some output classification schemes allowed only two discriminations, 0.0 to 0.5 and 0.5 to 1.0. However, although this type of strict binomial scheme may have given better results, we preferred to use a more stringent trinomial classification with predicted values near the middle of the range being considered inconclusive.

**Results**

Of the original parameters listed in the paper by Atherholt *et al.* (1998), only eight were required to predict *Cryptosporidium* peaks by neural network. Training models that had one parameter withheld, and comparing the prediction results on the validation set to that of a full parameter model, allowed parameters to be dropped. Eight parameters were found to be essential for prediction of *Cryptosporidium* peaks. (pH, *C. perfringens*, *E. coli*, faecal
coliform, turbidity, river flow, precipitation and total coliform). In the case of *Giardia* peak estimation, seven parameters were found to be essential (pH, *C. perfringens*, *E. coli*, faecal coliform, turbidity, river flow and alkalinity. As can be seen by comparing the significant input parameters, predictions for both types of encysted protozoa were dependent upon pH, *C. perfringens*, *E. coli*, faecal coliform, turbidity and river flow. This seemed reasonable since an increase in these parameters, with a decrease in pH, would coincide with events in the watershed that would bring more faecal material into the surface water. However, it would seem for *Cryptosporidium* that watershed events that flushed more soil into the river, as indicated by an increase in total coliform and averaged precipitation, were of more importance to peak events than for *Giardia*. This could be due to slight differences in the relative sources of the protozoa. Others have indicated that discharge from municipal sewage treatment plants tend to be a continuous source of encysted protozoa but to provide more high level loadings of *Giardia* than *Cryptosporidium* (Crockett and Haas, 1997). The sources in a watershed that have the greatest opportunity to input large numbers of *Cryptosporidium* are those of neonate cows and sheep. Therefore, precipitation as a transport vehicle for *Cryptosporidium* would be of greater importance while a rain event, significant enough to cause overflows from combined sewers and bypassing of the municipal waste treatment plant, would have much more impact upon *Giardia* concentrations and the river’s alkalinity. It should be noted that the training and validation sets of observations were identical for modelling attempts on both protozoa.

The best neural network architecture was one that doubled the number of internal nodes relative to the number of input nodes. That is, instead of having an architecture for the prediction of *Cryptosporidium* of eight input nodes with eight internal nodes and one output node (8-8-1), performance of the neural network modelling approach could be improved by using a 8-16-1 structure. This is clearly seen by analysis of Tables 1 and 2 where for both protozoa, the numbers of mis-prediction fall in the validation set when the number of internal nodes is increased. Of most significance was that the numbers of false negatives, times when a peak event would not have been predicted although it was in actuality occurring, fell to zero when the correct combination of input parameters and internal nodes was identified. While false positives still occurred, times when the numbers of encysted protozoa would not in actuality be above background levels as defined by this study, the goal of any type of early warning system would first be to avoid false negatives then to reduce false positives. For both protozoa, a neural network model could be trained to avoid false negatives in the

<table>
<thead>
<tr>
<th>Input parameter withheld</th>
<th>ANN architecture (input-internal-output)</th>
<th>Total observations (17)</th>
<th>Above background observations (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>8-8-1</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>None</td>
<td>8-16-1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td><em>C. perfringens</em></td>
<td>7-14-1</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td><em>E. coli</em></td>
<td>7-14-1</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>pH</td>
<td>7-14-1</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Total coliforms</td>
<td>7-14-1</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Faecal coliforms</td>
<td>7-14-1</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Precipitation</td>
<td>7-14-1</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Turbidity</td>
<td>7-14-1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>River flow</td>
<td>7-14-1</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>
validation set. These well trained models were then considered the standard against which other models were compared, models that had one of the vital input parameters left out. For both types of protozoa, it appeared that *C. perfringens* was a crucial input parameter. Neither model trained (data not shown) or predicted well on the validation set without including this microorganism. This finding was reasonable considering that the spores of *C. perfringens* originate from the same faecal sources and are very hardy in the environment. Coliforms are also of importance to predict protozoa peaks but total coliforms were found to be more important for *Cryptosporidium* than *Giardia*; however, both protozoa required faecal coliforms for optimum prediction. Turbidity appeared to be more important for predicting *Giardia* than for predicting peak *Cryptosporidium* events. This could be due to the fact that even a light rain, one that would not impact turbidity overmuch, could wash oocysts into surface waters. Whereas *Giardia* sewage overflows, that were a consistent and large source of cysts, would most likely be accompanied by larger shifts in turbidity from stronger precipitation events. The reliance of the *Cryptosporidium* model on precipitation as a critical input parameter as compared to the *Giardia* model supported this rationale. River flow was important for the prediction of protozoa but seasonal shifts in flow may have confounded how this variable impacted concentrations. It is thought that seasonal changes in environmental protozoa concentrations occur especially for *Cryptosporidium* that should peak when the population of neonate animals peaks. Therefore, it may be that river flow should be compared as a change from a moving average of previous values to better isolate storm events from seasonal changes in flow.

**Conclusions**

It is clear from this preliminary study that neural networks can learn to identify a unique set of water quality characteristics associated with peak concentrations of encysted protozoa. Neural networks were successfully applied to predict the peaks of *Cryptosporidium* and *Giardia* concentrations in the Delaware River that serves as a drinking water source. Results show that several surrogate indicators were strongly linked with protozoan presence with spores of *Clostridium perfringens* being one of the most important surrogates. Since each watershed is unique in the number and relative protozoan contributions of point and non-point sources, it may be difficult to build a single, multiple-indicator based, monitoring system for generalisation for all watersheds. The input parameters shown that were successful in predicting enhanced concentrations from the watershed studied in New Jersey may not be successful in other watersheds. Indeed, they may not be successful in other

<table>
<thead>
<tr>
<th>Input parameter withheld</th>
<th>ANN architecture (input-internal-output)</th>
<th>Total observations (17)</th>
<th>Above background observations (6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>7-7-1</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>None</td>
<td>7-14-1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><em>C. perfringens</em></td>
<td>6-12-1</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>Faecal coliforms</td>
<td>6-12-1</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Turbidity</td>
<td>6-12-1</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>River flow</td>
<td>6-12-1</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Alkalinity</td>
<td>6-12-1</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>pH</td>
<td>6-12-1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td><em>E. coli</em></td>
<td>6-12-1</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>
reaches of the same river system. However, the applicability of neural networks to learn the unique characteristic signature of multiple water quality surrogates for the prediction of peak protozoan concentrations on a case-by-case basis was an exciting finding and opened the door for individualised hazard warning and control expert systems for water utilities.

References