Estimating *Escherichia coli* loads in streams based on various physical, chemical, and biological factors

Dipankar Dwivedi,[1] Binayak P. Mohanty,[1] and Bruce J. Lesikar[1]

Received 20 November 2012; revised 4 April 2013; accepted 16 April 2013; published 28 May 2013.

[1] Microbes have been identified as a major contaminant of water resources. *Escherichia coli* is a commonly used indicator organism. It is well recognized that the fate of *E. coli* in surface water systems is governed by multiple physical, chemical, and biological factors. The aim of this work is to provide insight into the physical, chemical, and biological factors along with their interactions that are critical in the estimation of *E. coli* loads in surface streams. There are various models to predict *E. coli* loads in streams, but they tend to be system- or site-specific or overly complex without enhancing our understanding of these factors. Hence, based on available data, a Bayesian neural network (BNN) is presented for estimating *E. coli* loads based on physical, chemical, and biological factors in streams. The BNN has the dual advantage of overcoming the absence of quality data (with regard to consistency in data) and determination of mechanistic model parameters by employing a probabilistic framework. This study evaluates whether the BNN model can be an effective alternative tool to mechanistic models for *E. coli* load estimation in streams. For this purpose, a comparison with a traditional model (load estimator (LOADEST), U.S. Geological Survey) is conducted. The models are compared for estimated *E. coli* loads based on available water quality data in Plum Creek, Texas. All the model efficiency measures suggest that overall *E. coli* load estimations by the BNN model are better than the *E. coli* load estimations by the LOADEST model on all the three occasions (threefold cross validation). Thirteen factors were used for estimating *E. coli* loads with the exhaustive feature selection technique, which indicated that 6 of 13 factors are important for estimating *E. coli* loads. Physical factors included temperature and dissolved oxygen; chemical factors include phosphate and ammonia; and biological factors include suspended solids and chlorophyll. The results highlight that the LOADEST model estimates *E. coli* loads better in the smaller ranges, whereas the BNN model estimates *E. coli* loads better in the higher ranges. Hence, the BNN model can be used to design targeted monitoring programs and implement regulatory standards through total maximum daily load programs.

Citation: Dwivedi, D., B. P. Mohanty, and B. J. Lesikar (2013), Estimating *E. coli* loads in streams based on various physical, chemical, and biological factors, *Water Resour. Res.*, *49*, 2896–2906, doi:10.1002/wrcr.20265.

1. Introduction

[2] Microbes have been identified as a major contaminant (13.2% contamination caused by pathogenic microbes of total impaired water body segments) of water resources in the United States [*U.S. Environmental Protection Agency*, 2006]. Common bacterial waterborne pathogens include *Salmonella* sp., *Shigella* sp., few strains of *Escherichia coli*, *Pseudomonas aeruginosa*, *Aeromonas hydrophila*, *Mycobacteria*, *Helicobacter pylori*, and various others [*Fincher et al.*, 2009]. The most widely used indicator organisms are the enteric coliform bacteria, which are Gram-negative bacilli that belong to the family Enterobacteriaceae (e.g., *Klebsiella* spp., *Enterobacter* spp., *Citrobacter* spp., and *E. coli*) [*Hipsey et al.*, 2008; *Dorner et al.*, 2006; *Mead and Griffin*, 1998]. The indicator organisms are mostly harmless as compared to the pathogen(s) of concern. However, the indicator organisms are monitored due to the relative ease and lesser cost involved in their measurements. It is well established that the fate of *E. coli* in surface water systems is governed by multiple physical (e.g., temperature [*Flint*, 1987]), chemical (e.g., pH [*Sjogren and Gibson*, 1981], nutrients [*Lessard and Sieburth*, 1983], sulfate [*Robakis et al.*, 1983], and nitrate [*Noguchi et al.*, 1997]), and biological (chlorophyll [*Nevers and Whitman*, 2005]) factors. The relationship among these factors and *E. coli* loads gets complicated by flow rate [*Whitman et al.*, 2004; *McKergow and Davies-Colley*, 2009]. *Vidon et al.* [2008] have reported that *E. coli* loads are significantly higher at high flow than at low flow, whereas *McKergow and Davies-Colley* [2009] have

[1] Department of Biological and Agricultural Engineering, Texas A&M University, College Station, Texas, USA.

Corresponding author: B. P. Mohanty, Department of Biological and Agricultural Engineering, Texas A&M University, College Station, TX 77843-2117, USA. (bmohanty@tamu.edu)

0043-1397/13/10.1002/wrcr.20265

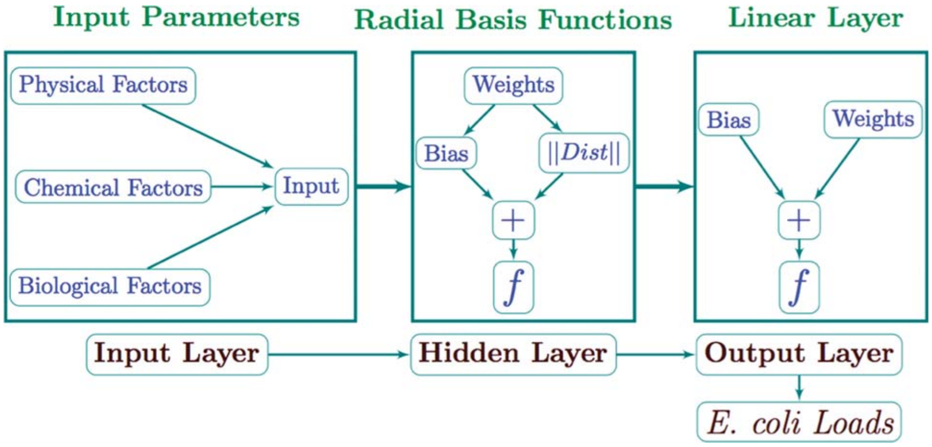

Figure 1. The graphical structure of BNN representing cause-and-effect relationship between system variables (water quality parameters) and the *E. coli* loads. The radial basis layer is the hidden layer, which uses the transfer function *f* (TPS); and the output layer is a linear layer, which uses the transfer function *f* (linear function). The transfer function *f* establishes a relationship between inputs and outputs; in case of estimation of *E. coli* loads, TPSs work better than other transfer functions (Gaussian or r^4 functions).

observed that *E. coli* peak loads always preceded discharge and turbidity peaks (which had similar timings). Therefore, *E. coli* evidently has a nonlinear relationship with the flow rate and the turbidity.

[3] It is important to develop an understanding of the relative importance of these physical, chemical, and biological factors in estimating the survival of *E. coli* in water bodies. However, a direct measurement of *E. coli* fate is not, in general, easy to implement. Therefore, the degree of impairment of a stream is assessed in terms of total maximum daily load (TMDL). Load duration curves are often used to estimate the reduction of contaminant loads in a watershed, especially in TMDL programs [*Babbar-Sebens and Karthikeyan*, 2009]. The load duration curves are measured using the instantaneous "load." The instantaneous "load" passing through a stream cross section is the product of the flow rate and the constituent concentration.

[4] Various models have been developed that use complex mechanistic and empirical relationships to predict the loads of *E. coli* in surface water systems, e.g., Soil and Water Assessment Tool [*Arnold and Fohrer*, 2005; *Pachepsky et al*., 2006], Hydrological Simulation Program – Fortran [*Benham et al*., 2006], and a watershed model developed by *Tian et al*. [2002]. However, overly complex mechanistic relationships and requirement of detailed descriptions of stream geometry and capacity, detailed information about sources within the watershed, sedimentation and resuspension characteristics, and bacteria die-off rates limit the utility of these models. Input parameter approximation and simplification in describing transport processes result in significant uncertainties in *E. coli* loads in streams. Other models have been developed that use statistical modeling framework to predict the loads of *E. coli* in surface water systems. For instance, *Nevers and Whitman* [2005] used regression modeling to determine *E. coli* using the wave height, lake chlorophyll, and turbidity for individual beaches of southern Lake Michigan. Furthermore, *Money et al*. [2009] estimated *E. coli* concentrations using turbidity, where *E. coli* data were not available, to assess fecal contamination along the Raritan River in New Jersey. Different models are relevant for different surface water environments, such as freshwater lakes and reservoirs [*Auer and Niehaus*, 1993; *Walker and Stedinger*, 1999; *Jin et al*., 2003; *Hipsey et al*., 2008], streams and rivers [*Wilkinson et al*., 1995; *Medema and Schijven*, 2001], and estuaries and coastal lagoons [*Steets and Holden*, 2003; *McCorquodale et al*., 2004]. It is also difficult for users to confidently implement these models since models tend to be system- or site specific.

[5] In comparison to these mechanistic and statistical models, a Bayesian neural network (BNN) provides a Bayesian modeling framework for estimating *E. coli* loads by utilizing routinely monitored flow rate and water quality data. The input data will comprise of water quality data (physical, chemical, and biological factors) that will provide a functional framework for the BNN. In case of sparse data sets, the Bayesian framework helps in representing input parameters as random variables emphasizing the statistical strength of the available data. Also, the uncertainty in input data sets is reflected through the probabilistic prediction of *E. coli* loads. The graphical structure of the BNN represents a cause-and-effect relationship between system variables (water quality data) and *E. coli* loads, as shown in Figure 1. One can use various basis functions in the formulation of the BNN such as multilayer perceptron or radial basis functions (RBFs). BNN models with RBFs have been used in this study as they have an ability to deal with sparse data sets and parameter overfitting [*Cilek and Yilmazer*, 2003].

[6] The specific objective of this study is to identify the key water quality factors for estimating the *E. coli* loads in streams. Based on identified water quality factors, *E. coli* loads will be estimated in streams along with characterization of possible uncertainties.

2. Study Area Description and Data Availability

[7] This study is conducted at a station (station ID: 12645; latitude 29°40′02″ and longitude 97°39′14″) in Plum Creek (Figure 2), which is monitored by the Texas

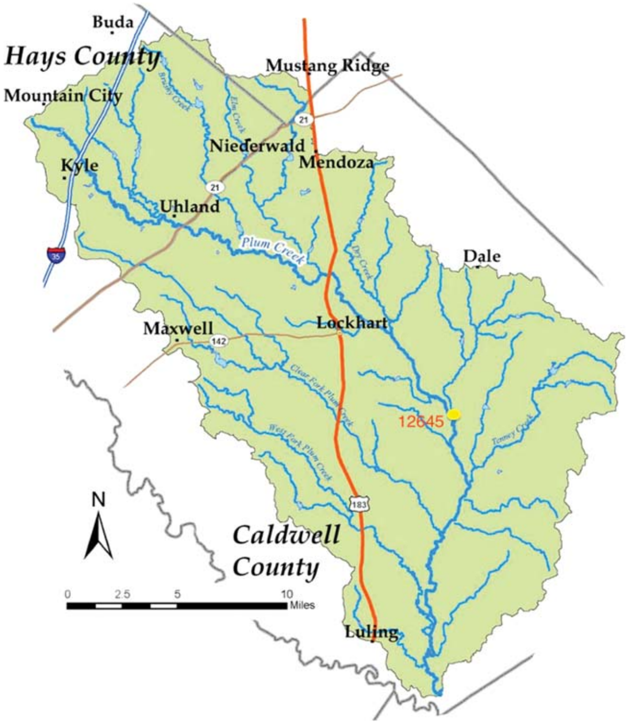

Figure 2. Map showing the Guadalupe River basin in east central Texas and the station ID 12645 in Plum Creek (map modified from http://www.gbra.org/CRP).

Commission on Environmental Quality (TCEQ). The Plum Creek watershed is a part of the Guadalupe River basin and is located in east central Texas. It surrounds a drainage area of 1028 km^2 in the counties of Hays, Caldwell, and Travis. According to the 2008 Texas Water Quality Inventory and 303(d) List of Impaired Water Bodies, Plum Creek is impaired for bacteria throughout the entire segment (http://www.gbra.org/CRP). Plum Creek is a shallow, intermittent fifth-order stream. It is 83 km long and joins the San Marcos River that in turn connects with the Guadalupe River. The watershed has several rapidly growing towns such as Lockhart, Kyle, and Luling. The watershed has a diversified land use from urban to agriculture and oil field activities. The watershed encompasses 38% rangeland, 17% pasture, 11% cultivated cropland, 18% forest, 8% developed land, 6% near riparian forest, and 2% open water and barren land. The landscape is characterized as rolling hills of pasture and cropland surrounded by scrub oak forest. Plum Creek lies in a semihumid subtropical climate zone and is heavily influenced by its proximity to the Gulf of Mexico (http://www.gbra.org/CRP).

[8] Two U.S. Geological Survey (USGS) gage stations are located on Plum Creek to monitor stream flows: one north of Lockhart (station 08172400) and one near Luling (station 08173000). Near Lockhart, periods of no flow have occurred almost every year on record. Southern reaches of Plum Creek, particularly south of Lockhart, are fed by a number of small springs and are usually perennial. Based on routine water quality sampling, the TCEQ initially listed portions of Plum Creek for bacteria impairment for contact recreation use in 2002. The possible sources of *E. coli* contamination in the creek are cows, livestock, wildlife, wastewater treatment plants, septic systems, and pet sources [*Teague et al*., 2009]. By 2004, bacterial contamination level in Plum Creek was elevated, and it was included in the list of impaired waters of Texas prohibiting wading and swimming. The *E. coli* criteria for designated use of a stream specified in water quality standards (e.g., recreational uses, irrigation, and navigation) require a geometric mean concentration of *E. coli* less than 126 cfu/100 mL of water with no sample exceeding 235 cfu/100 mL of water. *E. coli* and water quality data at the monitoring sites were available from October 1996 to December 2008 (http://www.gbra.org/CRP). Water quality data were collected monthly by the TCEQ. The available water quality data include 13 factors, wherein physical factors include turbidity (NTU), temperature (°C), conductivity (μmhos/cm), and dissolved oxygen (DO, mg/L); chemical factors include pH, phosphate (mg/L), nitrate-N (mg/L), chloride (mg/L), sulfate (mg/L), total hardness (mg/L), and ammonia (mg/L); and biological factors include suspended solids (SSs, mg/L) and chlorophyll (mg/m^3).

3. Methodology

[9] In order to identify the key water quality factors responsible for *E. coli* loads in streams, BNN models are run in conjunction with the exhaustive feature selection technique. We use the 13 physical, chemical, and biological factors described earlier. The exhaustive feature selection technique provides the best set of water quality factors for estimating *E. coli* loads. A principal component analysis (PCA) is also conducted to get insight into the relative importance of the factors identified by the exhaustive feature selection. These selected factors are subsequently utilized in estimating *E. coli* loads by the BNN model in Plum Creek. The BNN model results are also compared with the load estimator (LOADEST) [*Runkel et al*., 2004] model. For a better decision making, uncertainty analysis is also conducted. In the subsequent sections, we provide a description of BNN and LOADEST models, exhaustive feature selection, PCA, and uncertainty analysis.

3.1. Bayesian Neural Networks

[10] The application of the Bayesian learning paradigm to neural networks results in a flexible and powerful nonlinear modeling framework that can be used for regression, density estimation, prediction and classification supporting adaptive decision making, and accounting for uncertainties [*Andrieu et al*., 2001, *Reckhow*, 1999]. The regression of a target variable Y on an input set of covariates X given the data $D = \{(x_1, y_1), (x_2, y_2), \ldots\}$:

$$y_i = Mx_i + \xi_i, \tag{1}$$

where M is the model and ξ_i is independent and identically distributed error $\sim N(0, \sigma^2)$.

[11] The *E. coli* loads for each point (i) in time are estimated as follows:

$$(E.\ coli\ \text{Loads})_i = \text{BNN}\,(\text{DO}\,, \text{pH}\,, \ldots)_i + \xi_i. \tag{2}$$

[12] The central process of the Bayesian framework is the calculation of a probability distribution on the unknown

parameter (weight) vector **w**. Prior knowledge that we might have, say for small weights, is updated using the data. These posterior distributions are used in model predictions, with point forecasts given as expectations [*Holmes and Mallick*, 1998]:

$$E[Y|x,D] = \int m(x,w)p(w|D)\mathrm{d}w, \tag{3}$$

where $E[Y|x, D]$ represents the posterior probability of the parameters of the model $m(x, w)$ given the training data D.

[13] BNN generates a probability distribution of the layer weights, which is dependent on the given input data:

$$P(w|D) = \frac{P(Y|w,X)P(w)}{P(Y|X)}, \tag{4}$$

where $P(Y|X) = \int P(Y|w,X)P(w)\mathrm{d}w$ is the marginal distribution of Y, $P(w)$ is the prior distribution of weights, and $P(Y|w, X)$ is the likelihood function [*Gelman et al.*, 1995]. Artificial neural network combined with Monte Carlo Markov chain generates multiple samples from a continuous target density [*Bates and Campbell*, 2001]. A flat prior can be assumed here, as we do not have any concrete prior knowledge of weights [*Sims and Zha*, 1998].

[14] Predictive distribution of Y_{n+1} is given by

$$P(Y_{n+1}|x_{n+1},Y,X) = \int P(Y_{n+1}|x_{n+1},Y,w)P(w|Y,X)\mathrm{d}w, \tag{5}$$

where $n+1$ denotes the next realization.

[15] We considered RBF architecture in BNN, which has an ability of closely approximating any nonlinear multidimensional mapping [*Ciocoiu*, 2002]. A brief summary of RBF is provided later for completeness.

[16] RBF networks are one of the most commonly used types of feed forward networks. The feed forward neural network is most widely used to solve engineering problems. It is a simple nonlinear model that maps the input vector onto the output vector [*Lanouette et al.*, 1999].The architecture of a RBF network consists of three layers: an input layer, a hidden layer, and an output layer. The transformation from input space to hidden unit space is nonlinear, whereas transformation from hidden unit space to output space is linear [*Ciocoiu*, 2002]. During the training stage, a known set of input and output data pairs are delivered to the RBF network to select the centers and compute the output layer weights. The models have radial functions, where each basis is parameterized by a knot or position vector located in the d-dimensional covariate space x. The hidden layer provides a set of functions that constitute an arbitrary basis for the input patterns. The hidden units are known as radial centers and represented by the vectors (C_1, C_2, $\ldots$, C_h). Conventionally, there are as many basis functions (h) as data points to be approximated with the position vectors set to the data values. The model output $m(x)$ is given by a linear combination of the basis functions response and a low-order polynomial term:

$$m(x) = \sum_{i=1}^{N} w_i\varphi_i(||x-\mu_i||) + \sum_{m=0}^{p} a_m q_m(x), \tag{6}$$

where $\|.\|$ denotes a distance metric, usually Euclidean or Mahalanobis, and $q_m(x)$ represents a polynomial of degree m. The coefficients w and a are calculated by least squares where the constraint $\sum_{i=1}^{N} w_i q_m(x_i) = 0$ is imposed to ensure the uniqueness of the solution [*Holmes and Mallick*, 1998]. Different radial functions (e.g., Gaussian, quadratic, thin plate spline (TPS), inverse quadratic functions) are used for different problems. We used the TPS for estimating *E. coli* loads. The TPS is given as

$$\varphi(Z) = Z^2\log(Z), \tag{7}$$

where

$$Z = (||x-\mu_i||). \tag{8}$$

[17] RBF networks enlarge the dimensionality of the input data in order to increase the probability that originally nonlinearly separable classes become linearly separable (Cover's theorem) [*Ciocoiu*, 2002]. For modeling a system with limited experimental data, RBF has an advantage over the other techniques. One of the problems that may occur during neural network training is overfitting. A frequently used method for improving network generalization is to use an adequate-sized network, which is just large enough to provide an adequate fit [*Cilek and Yilmazer*, 2003]. Overfitting happens when the model has too many degrees of freedom, which is the result of including too many hidden neurons. The neurons in the hidden layer contain transfer functions whose outputs are inversely proportional to the distance from the center of the neuron. With small data sets used in this study, we ensured model accuracy by running multiple simulations by randomizing data sets. We split all the valid data into three randomly distributed groups. Three random sets are selected using "randperm" function in MATLAB. Initially, random split 1 is set aside for testing, while the models are parameterized on the basis of random splits 2 and 3. The fitted models are then used to test/predict *E. coli* loads by using input data from the random split 1. Next, random split 2 is set aside for testing, while random splits 1 and 3 are used for training. This pattern is also repeated for the random split 3. We use the same random splits (1–3) for estimating *E. coli* loads by using the LOADEST model.

3.2. Load Estimator

[18] LOADEST is a regression-based model for estimating constituent loads in streams and rivers [*Runkel et al.*, 2004]. Given a time series of streamflow and constituent concentration (*E. coli*), LOADEST facilitates users in developing a regression model for the estimation of constituent loads [*Cohn*, 2005]. Explanatory variables within the regression model include multiple functions of flow, time, and additional data variables. The developed regression model is then used to estimate loads over a user-specified time interval. Mean loads, standard errors, and 95% confidence intervals are also estimated on a monthly and/or seasonal basis. There are three statistical methods used for calibration and validation (estimation) of LOADEST, including adjusted maximum likelihood estimation (AMLE), maximum likelihood estimation (MLE), and least

absolute deviation (LAD). AMLE and MLE are appropriate when the calibration model errors (residuals) are normally distributed, whereas LAD is appropriate when model errors (residuals) are not normally distributed. In our case, calibration model errors are normally distributed, so we used AMLE for estimating *E. coli* loads. The detailed mathematical formulation of LOADEST is provided elsewhere [*Cohn*, 2005]. In general, total mass loading over an arbitrary time period, τ, is given by

$$L_\tau = \int_0^\tau QC\mathrm{d}t \tag{9}$$

$$\hat{L_\tau} = \Delta t \sum\nolimits_{t=1}^{\mathrm{NP}} \left(\hat{Q}C\right) = \Delta t \sum\nolimits_{t=1}^{\mathrm{NP}} \left(\hat{L}\right), \tag{10}$$

where C is the concentration [M/L^3], L is the total load [M], Q is the instantaneous stream flow [L^3/T], t is the time [T], and NP is the number of discrete points in time. The hats on Q, L_τ, and L denote the instantaneous values of the respective variables. *E. coli* loads estimated by the LOADEST model are compared with the *E. coli* loads estimated by the BNN model using the key water quality factors. The key water quality factors are identified using the exhaustive feature selection technique.

3.3. Exhaustive Feature Selection

[19] The BNN models are run multiple times with all possible combinations of the 13 water quality factors ($^{13}\mathrm{C}_1 + {}^{13}\mathrm{C}_2 + \ldots + {}^{13}\mathrm{C}_{13}$) for estimating *E. coli* loads. The exhaustive feature selection is a technique of selecting a subset of relevant features for building robust models. The brute-force feature selection algorithm is applied to exhaustively evaluate all possible combinations of the input features, and then the best subset is chosen. The exhaustive search's computational cost is prohibitively high, with a considerable danger of overfitting [*Moore and Lee*, 1994; *Skalak*, 1994]. Hence, for avoiding the overfitting, K-fold (threefold) cross validation is used in selecting the best subset. The aim of the feature selection is to choose a subset of the set of input features (physical, chemical, and biological factors) so that the subset can predict the output Y (*E. coli* loads) with accuracy akin to the performance of the whole input set X, and with a reduction of the computational cost. For conducting the exhaustive feature selection, the following steps are outlined: (1) Shuffle the data set and split into a training set of two third of the data and a test set of the remaining one third of the data. (2) Choose all possible combinations of various input variables. (3) Select each subset, and run the BNN model with leave-one-out cross validation. (4) Store the Nash-Sutcliffe efficiency (NSE) coefficients (see section 3.4) of each run. (5) Select the feature set which has minimum root-mean-square error of NSE threefold validation.

3.4. Model Performance

[20] We computed the NSE and normalized mean squared error (NMSE) as measures of the model performance.

[21] The NSE coefficient is given as

$$\mathrm{NSE} = 1 - \frac{\sum\nolimits_{t=1}^{T} \left(Q_0^t - Q_m^t\right)^2}{\sum\nolimits_{t=1}^{T} \left(Q_0^t - Q_0^E\right)^2} \tag{11}$$

[22] The NMSE is given as

$$\mathrm{NMSE} = \frac{\sum\nolimits_{n=1}^{N} \left(Q_0^t - Q_m^t\right)^2}{\mathrm{var}\left(Q_0^t\right)} \tag{12}$$

where Q_O^t is observed *E. coli* loads, Q_m^t is simulated *E. coli* loads at time t, Q_O^E is mean observed *E. coli* loads, and $\mathrm{var}\left(Q_O^t\right)$ denotes the variance of all the observed *E. coli* loads. NSE can range from $-\infty$ to 1. An efficiency of 1 (NSE = 1) corresponds to a perfect match of simulated values to the observed data. An efficiency of 0 (NSE = 0) demonstrates that the model predictions are as accurate as the mean of the observed data. In essence, closer the efficiency of the model is to 1, the more accurate is the model. The NMSE of 0 indicates that the model predictions are perfect. The lower the NMSE, the better is the model performance.

[23] The exhaustive feature selection technique in conjunction with the BNN model rendered the best set of factors. In order to assess the relative importance of these factors, PCA is done.

3.5. Principal Component Analysis

[24] PCA is a multivariate statistical technique. The transformed features have a descriptive power that is more ordered than the original features. PCA has been applied in describing various aspects of streamflow regimes [*Olden and Poff*, 2003], understanding the spatial and temporal changes in water quality [*Bengraïne and Marhaba*, 2003], and determination of dominant biogeochemical processes in a contaminated aquifer [*Baez-Cazull et al*., 2008]. In this study, PCA is used to identify major factors among water quality data that can explain most of the variation of *E. coli* loads.

[25] PCA is an orthogonal linear transformation of the data (e.g., water quality data) to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first quadrant. The principal axis method is used to extract the components, followed by a varimax (orthogonal) rotation with Kaiser normalization. A detailed description, of how the principal components (PCs) are calculated, is provided elsewhere [*Jolliffe*, 2002].

3.6. Uncertainty Analysis

[26] Monte Carlo based statistical techniques, resampling with replacement ("bootstrapping") [*Robert and Casella*, 1999], are implemented to estimate the statistical uncertainty in predictions by the BNN and LOADEST models. To explore the uncertainty in the BNN predictions, 10,000 realizations of *E. coli* loads are investigated. Bayesian networks are probabilistic models that combine prior distributions of uncertainty with data to yield an updated (posterior) set of distributions [*Helton and Oberkampf*, 2004]. Therefore, inputs are integrated over the weight space of the posterior probability distribution for finding the outputs (i.e., *E. coli* loads) of the networks.

[27] The probability distribution of each output as a random variable is plotted utilizing the kernel density (Parzen window) estimation, which is a nonparametric method [*Silverman*, 1986]. If $x_1, x_2, \ldots, x_N$ are samples drawn from the density function of a random variable, then the kernel density approximation of its probability density function is given as

$$\hat{f}_h(x) = \frac{1}{N}\sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right), \quad (13)$$

where K is some kernel and h is a smoothing parameter called the bandwidth. Here a Gaussian kernel is chosen with mean zero and unit variance:

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-(x^2/2)}. \quad (14)$$

4. Results and Discussion

[28] The results presented here provide insight into the different physical, chemical, and biological factors that are critical in the estimation of *E. coli* loads in surface streams. The exhaustive feature selection in conjunction with the BNN model (Figure 1) identified the best combination of input variables for estimating *E. coli* loads. Out of the 13 water quality factors, exhaustive feature selection identified 6 key variables: SS, phosphate, temperature, DO, ammonia, and chlorophyll. In the following section, we will focus on these key variables and their relative importance. Subsequently, we utilize these six key variables using the BNN model for estimating *E. coli* loads in Plum Creek.

4.1. Identification of the Key Factors Responsible for the *E. coli* Loads in Plum Creek

[29] The exhaustive feature selection identified six factors in estimating *E. coli* loads in Plum Creek, namely, SS, phosphate, temperature, DO, ammonia, and chlorophyll. To investigate the relative importance of the key factors for *E. coli* loads in streams, a PCA was performed as shown in Figures 3 and 4. The PCA explored the relationship among water quality factors such as SS, phosphate, temperature, DO, ammonia, and chlorophyll. The first two components explain 60.0% of the variance; component 1 and component 2 account for 35.6% and 24.4% of the variance, respectively (Figure 4). The first PC captures the variance of DO and temperature. The second PC captures the variance of SS, phosphate, ammonia, and chlorophyll. The PCA biplot (Figure 4) illustrates a visual interpretation of the factor loadings that result from a bicluster system of variables projected onto the first and second PC axes. The biplot tells about the relative positions of the factors, and the angles between the factors give approximate estimates of the correlation among factors; small angles between projected axes imply a high correlation. The direction of axes gives the sign of correlation among factors displayed on the biplot [*Jolliffe*, 2002].

[30] Turning now to the interpretation of the PCs in the present work, the six factors can be divided into three groups. Group 1 includes temperature and DO (physical factors); group 2 includes phosphate and ammonia (chemical factors); and group 3 includes the SS and the chlorophyll (biological factors). The central idea of this classification is based on the fact that groups of variables often move together, and more than one factor measures the same driving force. The first PC clearly measured physical factors, as DO and temperature have the maximum loadings (Figure 4b); moreover, they also have a high negative correlation with each other (almost 180° separated in biplot; Figure 4a). Therefore, DO and temperature were classified as group 1. The second PC accounted for the chemical and biological factors (Figure 4c). Since phosphate and ammonia are the dominant factors on the second PC, they also have a high positive correlation with each other (almost overlapping in biplot; Figure 4a). For this reason, they were classified as group 2. Similarly, the third PC also accounted for biological factors, as SS and chlorophyll have a medium positive correlation with each other (a small angle between them in biplot Figure 4a). Hence, they were grouped together.

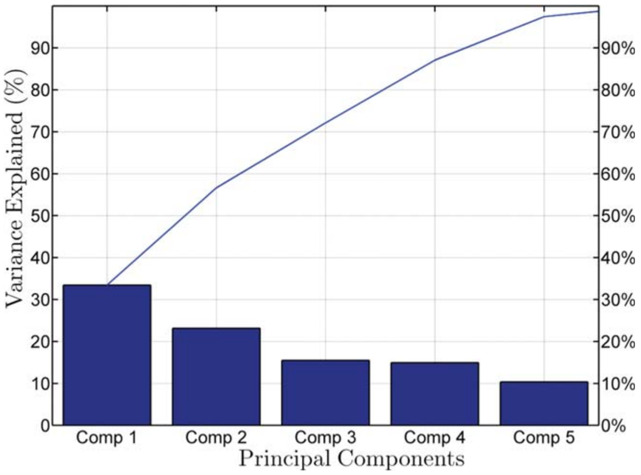

Figure 3. Pareto diagram of PCs shows the percentage explained by each component. The first three components explain 70% of variation in the data set.

[31] The biological tolerance of *E. coli* to different physical, chemical, and biological factors has been well studied, albeit mostly in the laboratory. It has been observed that *E. coli* are sensitive to changes in temperature [*Maeda et al.*, 1976; *Berg*, 2004]. The rate of die-off depends on temperature [*Flint*, 1987]. Moreover, *E. coli* are anaerobic bacteria, and thus, *E. coli* also responds to oxygen gradient. The majority of *E. coli* cannot live in oxygen rich environment [*Berg*, 2004]. This clearly explains the selection of DO and temperature as important physical factors in estimating *E. coli* loads in our study and their negative correlation. Temperature affects positively, whereas DO affects negatively in estimating *E. coli* loads by the BNN model. In the biplot, approximately 180° separation of temperature and DO corroborates this behavior of *E. coli* (Figure 4a).

[32] Phosphate and ammonia are also found to be important factors in the estimation of *E. coli* loads by the BNN model. This is because phosphate and ammonia act as nutrients or substrates, and the presence of nutrients

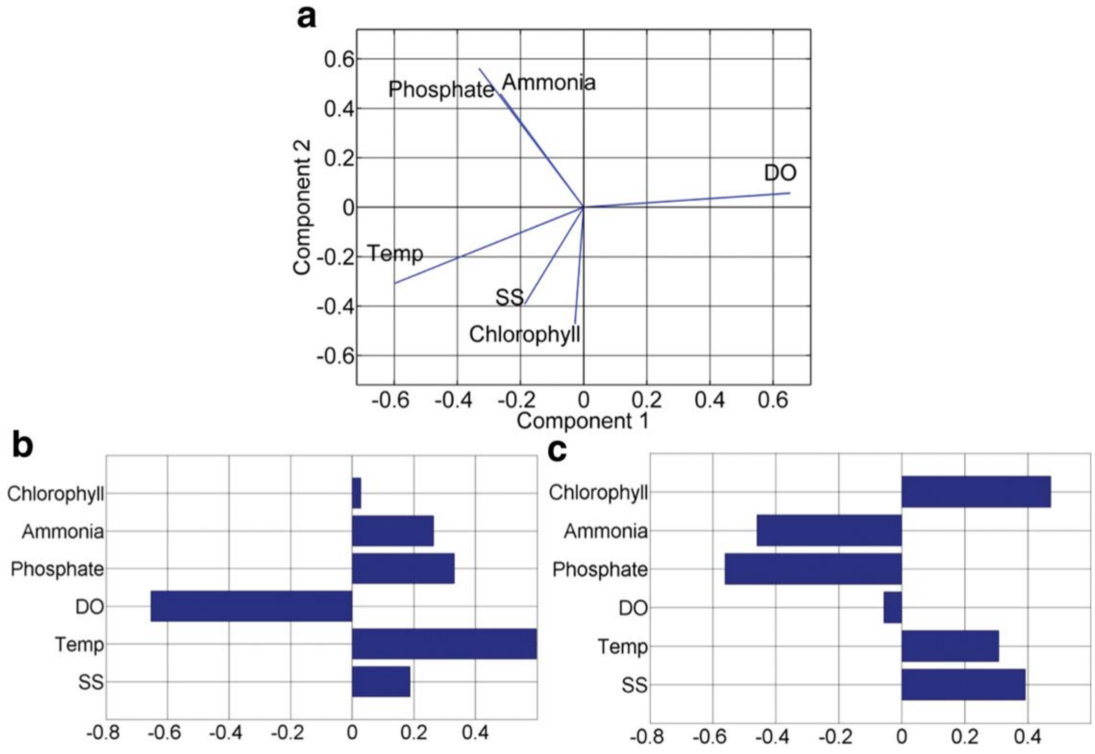

Figure 4. (a) Biplot of PCA is plotted by projecting the first PC against the second PC. Factor loadings on the (b) first and (c) second PCs reflect the relative importance of each factor.

increases *E. coli* concentrations in streams [*Van der Steen et al*., 2000]. These nutrients have significant positive correlation and therefore signify the importance of chemical factors on *E. coli* loads.

[33] In our study, SS and chlorophyll are also important factors in the estimation of *E. coli* loads by the BNN model. In literature, there is evidence to suggest that high concentrations of chlorophyll and suspended sediments are associated with high *E. coli* concentrations [*Nevers and Whitman*, 2005]. However, *Money et al*. [2009] examined the relationship between turbidity and *E. coli* and found a significant correlation between both the parameters. Turbidity indicates high volumes of suspended sediments. SS and chlorophyll correspond to the biological factors, as they are sources of organic carbon [*de Jonge*, 1980]. These biological factors were measured by the second PC.

[34] It should be noted that the sign of any PC is completely arbitrary. If every coefficient in a PC has its sign reversed, the variance is unchanged and so is the orthogonality [*Jolliffe*, 2002]. Therefore, the biplot and loadings only show the relative importance of the factors, they do not demonstrate if a factor is positively or negatively affecting the *E. coli* loads. However, the biplot exhibits how each factor can affect the *E. coli* loads. For example, it is evident from Figure 4a that all the factors on the left side of the plot (phosphate, ammonia, temperature, SS, and chlorophyll) are positively associated with *E. coli* loads, whereas the only factor on the right-hand side of the plot is DO, and it is negatively associated with *E. coli* loads. This graphic examination further substantiates our findings.

4.2. Estimation of *E. coli* Loads

[35] In this section, we discuss the discrepancy between simulated and observed *E. coli* loads using the BNN model (using the six key variables) and compare its performance to the LOADEST model. Figures 5 and 6 show measured and simulated *E. coli* loads in Plum creek using the LOADEST and BNN models, respectively, for three random splits. Table 1 shows the measures of the models' performance. A threefold cross validation results show that both modeling approaches (BNN and LOADEST) reproduce observed *E. coli* loads reasonably well, with all NSE values greater than or equal to 0.39 and all NMSE values smaller than or equal to 0.59 (Table 1). However, the BNN is able to estimate *E. coli* loads better in all the three random splits (Table 1). The uncertainty bands (Figures 5 and 6) show that the BNN is also able to capture higher *E. coli* loads more accurately than the LOADEST model. This is expected because the BNN model provides more flexible choices for the functional dependence in estimating *E. coli* loads based on physical, chemical, and biological factors (e.g., SS, phosphate, temperature, DO, ammonia, and chlorophyll), whereas the LOADEST model uses only the *E. coli* and flow data.

[36] Figure 7 shows the cumulative distribution functions of the observed, BNN simulated (all the three random splits), and LOADEST simulated (all the three random splits) *E. coli* loads in Plum Creek. Region A signifies the smaller *E. coli* loads (smaller than 0.5×10^{11} cfu/d), which are better estimated by the LOADEST model. The BNN model is underestimating the *E. coli* loads in this region. Region C encompasses the higher *E. coli* loads (greater than 1.5×10^{11} cfu/d), which are better estimated by the

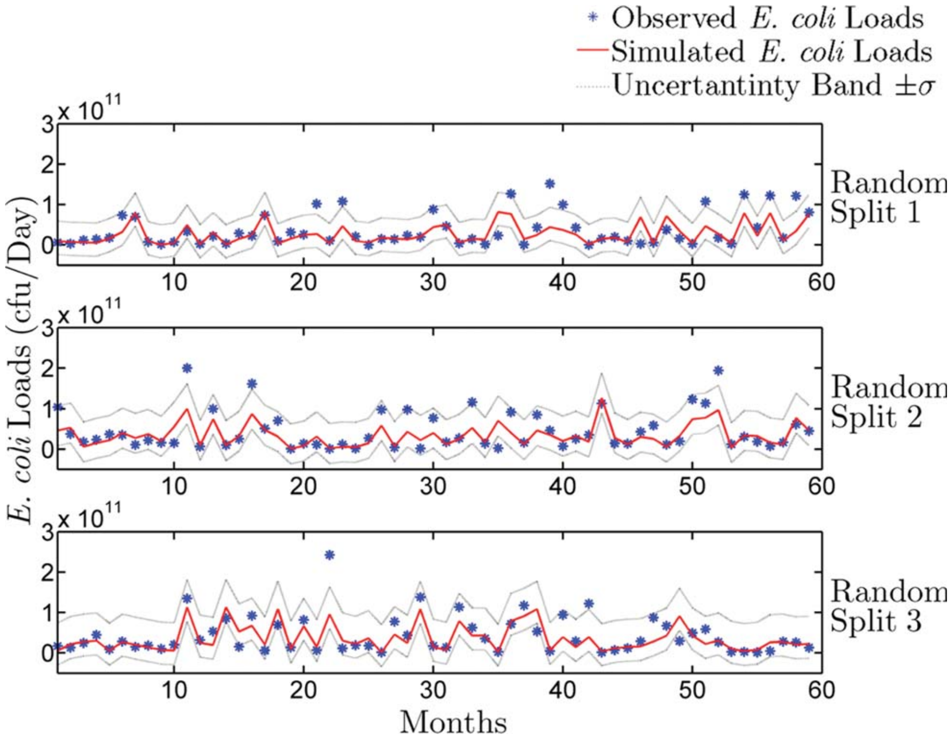

Figure 5. Measured and simulated loads of *E. coli* by the LOADEST model in Plum Creek are presented here for the three random splits. These random splits were used for threefold cross validation of the BNN model.

BNN model. For best management practices, it is essential to be able to estimate higher *E. coli* loads, and the BNN model is able to estimate values with greater accuracy in this range. Region B (0.5 × 10^{11} to 1.5 × 10^{11} cfu/d) constitutes the region with medium loads between regions A and C, and where both the LOADEST and BNN models

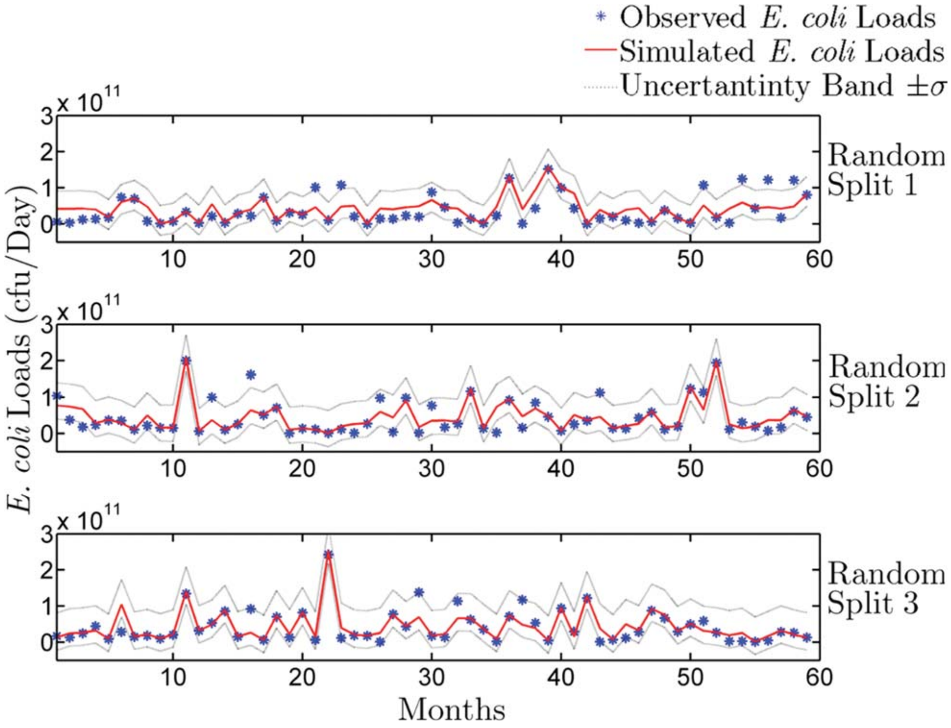

Figure 6. Measured and simulated loads of *E. coli* by the BNN model in Plum Creek are presented here. The simulations were tested by threefold cross validation.

Table 1. NSE and NMSE of Estimated *E. coli* Loads by BNN and LOADEST Models and Observed *E. coli* Loads in Plum Creek

Random Splits	Models	NSE	NMSE
Random split 1	BNN	0.48	0.51
	LOADEST	0.39	0.59
Random split 2	BNN	0.69	0.30
	LOADEST	0.55	0.44
Random split 3	BNN	0.75	0.23
	LOADEST	0.52	0.46

are overestimating the *E. coli* loads. However, the BNN model is closer to the observed values than the LOADEST model in this region.

4.3. Sensitivity Analysis

[37] In this study, the objective of the sensitivity analysis is to demonstrate the relative response (*E. coli* loads) of the BNN model for each physical (temperature and DO), chemical (phosphate and ammonia), and biological (SS and chlorophyll) factor. One-factor-at-a-time (OFAT) approach has been deemed appropriate for evaluating the sensitivity of different explanatory variables in the neural network models [e.g., *Xie et al*., 2007; *Delen et al*., 2006]. The OFAT approach involves perturbing each factor individually within a reasonable interval (±20%) and keeping the rest of the factors constant at their baseline values. The effect of perturbation of a single factor is quantified by recording the corresponding variation in the BNN output using NMSE and percentage change in the *E. coli* loads. NMSE values are calculated by using residuals between *E. coli* loads estimated by original variables and *E. coli* loads estimated by perturbing each factor. A higher NMSE value implies a higher sensitivity to the factor under consideration. Correspondingly, a higher percentage change in the *E. coli* loads means a higher sensitivity to that factor. Table 2 lists the relative sensitivity of each factor for all the three random splits. This ranking suggests that each factor shows comparable sensitivity; however, DO and SS show higher sensitivity as compared to other factors (temperature, phosphate, ammonia, and chlorophyll) in estimating *E. coli* loads.

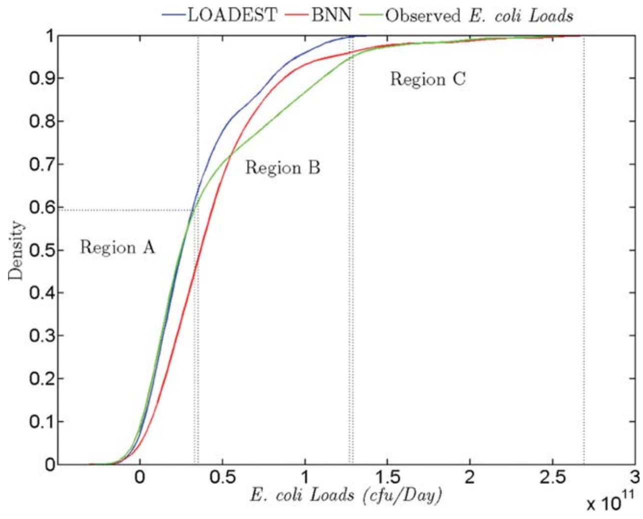

Figure 7. Cumulative density functions of the observed *E. coli* loads, estimated *E. coli* loads by the BNN, and estimated *E. coli* loads by the LOADEST are presented here. There are three regions in this figure. The region A signifies that the LOADEST model is able to estimate *E. coli* loads better. In regions B and C, the BNN model predicts better than the LOADEST model.

Table 2. NMSE and Percentage Change of Estimated *E. coli* Loads Due to Perturbation of Each Factor (±20%) One-At-A-Time and Estimated *E. coli* Loads While Keeping Other Factors Constant at Their Baseline Values by the BNN Model

Factors	Variation	Random Split 1	Random Split 2	Random Split 3	% Change in *E. coli* Loads (Range)
Temperature	Lower	0.22	0.31	0.32	−5 to 5
	Upper	0.27	0.34	0.27	
DO	Lower	0.43	0.45	0.56	−25 to 25
	Upper	0.35	0.37	0.32	
Phosphate	Lower	0.18	0.10	0.21	−2 to 12
	Upper	0.21	0.23	0.22	
Ammonia	Lower	0.15	0.21	0.22	−2 to 12
	Upper	0.14	0.11	0.12	
SS	Lower	0.39	0.37	0.43	−20 to 20
	Upper	0.51	0.43	0.47	
Chlorophyll	Lower	0.23	0.14	0.17	−2 to 5
	Upper	0.22	0.24	0.18	

4.4. Uncertainty Analysis

[38] Uncertainty analysis is conducted to further compare the performance of BNN and LOADEST models in estimating *E. coli* loads in Plum Creek. The uncertainty bands (±σ with 95% confidence) computed using bootstrap samples show that there is more uncertainty for larger loads than smaller loads (Figures 5 and 6). There is evidence that uncertainties of discrete *E. coli* samples are greater than 30%, while the uncertainties in storm water flow measurements are greater than 97% [*McCarthy et al*., 2008]. Therefore, *E. coli* loads will have more uncertainty due to storm events. As high *E. coli* loads are often associated with storm events, the upper limit of the uncertainty band is also wider for higher loads. These uncertainties in the inputs propagate into larger uncertainties in the output.

[39] Figure 8 shows six *E. coli* loads estimated by the BNN model (low and high *E. coli* loads for each random split), and the probability distribution functions (PDFs) of 10,000 realizations for each *E. coli* loads were plotted. As stated previously, BNNs use a range of weight sets instead of a single set. Each weight gives a realization of *E. coli* loads. The final predicted *E. coli* loads were generated from the average of 10,000 such realizations. The center of mass of a PDF shows the mean of the prediction and spread around the mean shows the uncertainty. It is clear from Figure 8 that *E. coli* loads, estimated by the BNN model, were closer to the centers of the PDFs with high density values (four of them are >0.4). It should be noted that the BNN model estimates lower *E. coli* loads with a small bias (observed *E. coli* loads falling close to the center of mass of the PDFs) and higher *E. coli* loads with a large bias (observed *E. coli* loads falling on the tails of the PDFs); however, the performance of the BNN model is better than the LOADEST model for estimating higher *E. coli* loads.

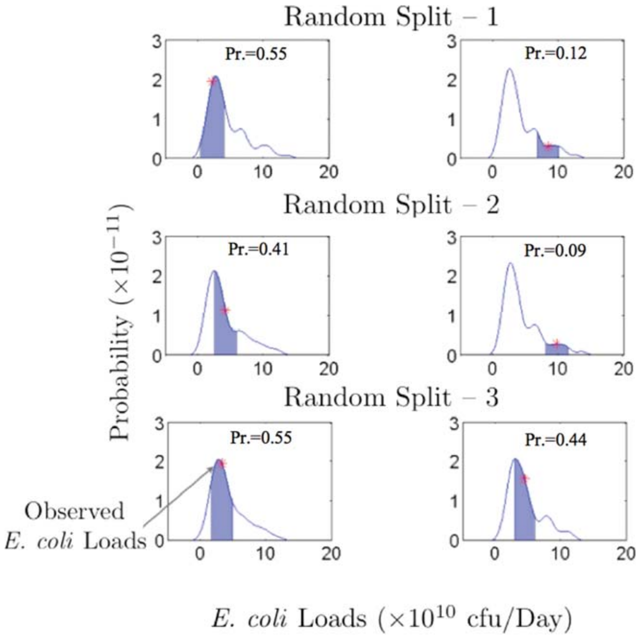

Figure 8. Probability distributions of (left) low and (right) high *E. coli* loads of each random split by the BNN model. PDFs show that higher loads are associated with multimodality. The blue shaded area (Pr.) represents the probability of observed *E. coli* loads with in ±10% uncertainty bands.

The large variance in the PDFs is due to various uncertainties, which stem mainly from (1) the uncertainties in input data (e.g., flow rate and water quality data) and (2) uncertainties in data used for calibration (e.g., *E. coli* loads). Input data (flow rate and water quality data) and *E. coli* loads have large inherent uncertainties, and these uncertainties cannot be removed from the model predictions in the existing data. However, the advent of newer technologies and careful data collection may help in minimizing these uncertainties in the future. The other source of uncertainties is from model parameters (weights and biases). These uncertainties are related to the fact that a small bias in the estimation, using a neural network with a training set of fixed size, can only be achieved with a large variance [*Geman et al*., 1992; *Haykin*, 1996]. This dilemma can be avoided if the training set is made very large, but the total amount of data is limited in our case. However, a possibility of making training sets larger can be plausible in the future.

5. Conclusions

[40] This study provides a BNN model for *E. coli* prediction in streams. A significant contribution of this paper is in identifying six key variables from a selection of physical, chemical, and biological factors that influence *E. coli* loads in surface streams. An exhaustive feature selection technique used in conjunction with BNN and the PCA indicated the importance and correlation among these six variables. Physical factors included temperature and DO; chemical factors included phosphate and ammonia; and biological factors included SS and chlorophyll. The sensitivity analysis was conducted on these factors which demonstrated all the six factors to be sensitive and DO and SS to be the most sensitive with respect to estimating *E. coli* loads.

[41] The BNN model was then run using these six factors, and a comparison with a traditional model (LOADEST) developed by the USGS was also conducted. The inherent differences between the models are the calibration procedures using statistical (LOADEST) versus probabilistic (BNN) framework. The models were compared for the estimation of *E. coli* loads based on available water quality data using NSE and NMSE in threefold cross validation. All the efficiency measures suggest that estimation of *E. coli* loads by the BNN model was better than the LOADEST model on all the occasions during threefold cross validation. The results also highlight that the LOADEST model estimates *E. coli* loads better in the smaller ranges, whereas the BNN model estimates *E. coli* loads better in the higher ranges, as well. Hence, the BNN model can be useful to decision maker and environmental managers to design targeted monitoring programs and establishing regulatory control such as TMDL programs. An uncertainty analysis is also used to compare the predictive powers of the two models. These results suggest that more uncertainty is associated with larger *E. coli* loads, and signify that the major source of uncertainty comes from storm events associated with *E. coli* loads.

[42] **Acknowledgments.** This project was supported by EPA 319(h) grant for TMDL in Texas streams. This research was also partly supported by National Science Foundation (grant EAR 0635961) and National Institute of Environmental Health Sciences (grant 5R01ES015634). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

References

Andrieu, C., N. de Freitas, and A. Doucet (2001), Robust full Bayesian learning for radial basis networks, *Neural Comput.*, *13*, 2359–2407.

Arnold, J. G., and N. Fohrer (2005), SWAT2000: current capabilities and research opportunities in applied watershed modelling, *Hydrol. Processes*, *19*(3), 563–572.

Auer, M. T., and S. L. Niehaus (1993), Modeling fecal coliform bacteria: 1. Field and laboratory determination of loss kinetics, *Water Res.*, *27*, 693–701.

Babbar-Sebens, M., and R. Karthikeyan (2009), Consideration of sample size for estimating contaminant load reductions using load duration curves, *J. Hydrol.*, *372*(1–4), 118–123.

Baez-Cazull, S. E., J. T. McGuire, I. M. Cozzarelli, and M. A. Voytek (2008), Determination of dominant biogeochemical processes in a contaminated aquifer-wetland system using multivariate statistical analyses, *J. Environ. Qual.*, *37*(1), 30–46.

Bates, B. C., and E. P. Campbell (2001), A Markov Chain Monte Carlo scheme for parameter estimation and inference in conceptual rainfall–runoff modeling, *Water Resour. Res.*, *37*(4), 937–947.

Bengraïne, K., and T. F. Marhaba (2003), Using principal component analysis to monitor spatial and temporal changes in water quality, *J. Hazard. Mater. B*, *100*, 179–195.

Benham, B. L., et al. (2006), Modeling bacteria fate and transport in watersheds to support TMDLs, *Trans. ASABE*, *49*, 987–1002.

Berg, H. C. (2004), *E. coli in Motion*, pp. 10–30, Springer-Verlag, New York, N.Y.

Cilek, E. C., and B. Z. Yilmazer (2003), Effects of hydrodynamic parameters on entrainment and flotation performance, *Miner. Eng.*, *16*, 745–756.

Ciocoiu, I. B. (2002), Hybrid feed forward neural networks for solving classification problems, *Neural Process. Lett.*, *16*, 81–91.

Cohn, T. A. (2005), Estimating contaminant loads in rivers: An application of adjusted maximum likelihood to type 1 censored data, *Water Resour. Res.*, *41*, W07003, doi:10.1029/2004WR003833.

de Jonge, V. N. (1980), Fluctuations in the organic carbon to chlorophyll a ratios for estuarine benthic diatom populations, *Mar. Ecol. Prog. Ser.*, *2*, 345–353.

Delen, D., R. Sharda, and M. Bessonov (2006), Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks, *Accid. Anal. Prev.*, *38*(3), 434–444.

Dorner, S. M., W. B. Anderson, R. M. Slawson, N. Kouwen, and P. M. Huck (2006), Hydrologic modeling of pathogen fate and transport, *Environ. Sci. Technol.*, *40*, 4746–4753.

Fincher, L. M., C. D. Parker, and C. P. Chauret (2009), Occurrence and antibiotic resistance of *Escherichia coli* O157:H7 in a watershed in north-central Indiana, *J. Environ. Qual.*, *38*,997–1004, doi:10.2134/jeq2008.0077.

Flint, K. P. (1987), The long-term survival of *Escherichia coli* in river water, *J. Appl. Bacteriol.*, *63*, 261–270.

Gelman, A., B. J. Carlin, H. S. Stern, and D. B. Rubin (1995), *Bayesian Data Analysis*, pp. 3–14, CRC Press, London, U. K.

Geman, S., E. Bienenstock, and R. Doursat (1992), Neural networks and the bias variance dilemma, *Neural Comput.*, *4*, 1–58.

Haykin, S. (1996), Neural networks expand SP's horizons, *IEEE Signal Process. Mag.*, *13*, 24–49.

Helton, J. C., and W. L. Oberkampf (2004), Alternative representations of epistemic uncertainty, *Reliab. Eng. Syst. Saf.*, *85*, 1–10.

Hipsey, M. R., J. P. Antenucci, and J. D. Brookes (2008), A generic, process-based model of microbial pollution in aquatic systems, *Water Resour. Res.*, *44*, W07408, doi:10.1029/2007WR006395.

Holmes, C. C., and B. K. Mallick (1998), Bayesian radial basis functions of variable dimension, *Neural Comput.*, *10*, 1217–1233.

Jin, G., A. J. Englande, and A. Liu (2003), A preliminary study on coastal water quality monitoring and modeling, *J. Environ. Sci. Health*, *A38*, 493–509.

Jolliffe, I. T. (2002), *Principal Component Analysis*, pp. 63–130, Springer-Verlag, New York, N.Y.

Lanouette, R., J. Thibault, and J. L. Valade (1999), Process modeling with neural networks using small experimental data sets, *Comput. Chem. Eng.*, *23*, 1167–1176.

Lessard, E. J., and J. M. Sieburth (1983), Survival of natural sewage populations of enteric bacteria in diffusion and batch chambers in the marine-environment, *Appl. Environ. Microbiol.*, *45*, 950–959.

Maeda, K., Y. Imae, J. I. Shioi, and F. Oosawa (1976), Effect of temperature on motility and chemotaxis of *Escherichia coli*, *J. Bacteriol.*, *127*, 1039–1046.

McCarthy, D. T., A. Deletic, V. G. Mitchell, and C. Diaper (2008), Uncertainties in storm water *E. coli* levels, *Water Res.*, *42*(6–7), 1812–1824.

McCorquodale, J. A., I. Georgiou, S. Carnelos, and A. J. Englande (2004), Modeling coliforms in storm water plumes, *J. Environ. Eng. Sci.*, *3*, 419–431.

McKergow, L. A., and R. J. Davies-Colley (2009), Stormflow dynamics and loads of *Escherichia coli* in a large mixed land use catchment, *Hydrol. Processes*, *24*(3), 276–289, doi:10.1002/hyp.7480.

Mead, P. S., and P. M. Griffin (1998), *Escherichia coli* O157: H7, *Lancet*, *352*, 1207–1212.

Medema, G. J., and J. F. Schijven (2001), Modelling the sewage discharge and dispersion of Cryptosporidium and Giardia in surface water, *Water Res.*, *35*, 4307–4316.

Money, E. S., G. P. Carter, and M. L. Serre (2009), Modern space/time geostatistics using river distances: Data integration of turbidity and *E. coli* measurements to assess fecal contamination along the Raritan River in New Jersey, *Environ. Sci. Technol.*, *43*, 3736–3742, doi:10.1021/es803236j.

Moore, A. W., and M. S. Lee (1994), Efficient algorithms for minimizing cross validation error, in Proceedings of the 11th International Conference on Machine Learning, ML-94, Morgan Kaufmann, New Brunswick, N.J.

Nevers, B. M., and R. L. Whitman (2005), Nowcast modeling of *Escherichia coli* concentrations at multiple urban beaches of southern Lake Michigan, *Water Res.*, *39(20)*, 5250–5260, doi.org/10.1016/j.watres.

Noguchi, K., H. Nakajima, and R. Aono (1997) Effects of oxygen and nitrate on growth of *Escherichia coli* and *Pseudomonas aeruginosa* in the presence of organic solvents, *Extremophiles*, *1*, 193–198.

Olden, J. D., and N. L. Poff (2003), Redundancy and the choice of hydrologic indices for characterizing streamflow regimes, *River Res. Appl.*, *19*, 101–121.

Pachepsky, Ya. A., A. M. Sadeghi, S. A. Bradford, D. R. Shelton, A. K. Guber, and T. H. Dao (2006), Transport and fate of manure-borne pathogens: Modeling perspective, *Agric. Water Manage.*, *86*, 81–92.

Reckhow, K. H. (1999), Water quality prediction and probability network models, *Can. J. Fish. Aquat. Sci.*, *56*, 1150–1158.

Robakis, N., Y. Cenatiempo, L. Meza-Basso, N. Brot, and H. Weissbach (1983), A coupled DNA-directed in vitro system to study gene expression based on di- and tripeptide formation, *Methods Enzymol.*, *101*, 690–706.

Robert, C. P., and G. Casella (1999), *Monte Carlo Statistical Methods*, pp. 32–35, Springer, New York.

Runkel, R., C. G. Crawford, and T. A. Cohn (2004), Load Estimator (LOADEST): A Fortran Program for Estimating Constituent Loads in Streams and Rivers, Book 4, chap. A5, U.S. Geol. Surv. Tech. Methods, Reston, Va.

Sims, C. A., and T. Zha (1998), Bayesian methods for dynamic multivariate models, *Int. Econ. Rev.*, *39*, 949–968.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, U.K.

Sjogren, R. E., and M. J. Gibson (1981), Bacterial survival in a dilute environment, *Appl. Environ. Microbiol.*, *41*, 1331–1336.

Skalak, D. B. (1994), Prototype and feature selection by sampling and random mutation hill climbing algorithms, in Proceedings of the 11th International Conference on Machine Learning, ML-94, Morgan Kaufmann, New Brunswick, NJ.

Steets, B. M., and P. A. Holden (2003), A mechanistic model of runoff associated fecal coliform fate and transport through a coastal lagoon, *Water Res.*, *37*, 589–608.

Teague, A., R. Karthikeyan, M. Babbar-Sebens, R. Srinivasan, and R. A. Persyn (2009), Spatially explicit load enrichment calculation tool to identify potential *E. coli* sources in watersheds, *Trans. ASABE*, *52*(4), 1109–1120.

Tian, Y. Q., P. Gonga, J. D. Radkeb, and J. Scarborough (2002), Spatial and temporal modeling of microbial contaminants on grazing farmland, *J. Environ. Qual.*, *31*, 860–869.

U.S. Environmental Protection Agency (2006), Available at http://www.epa.gov/volunteer/stream/vms50.html [accessed 26 June 2009] and http://www.gbra.org/CRP/Sites [accessed 12 May 2009].

Van der Steen, P., A. Brenner, Y. Shabtai, and G. Oron (2000), Improved fecal coliform decay in integrated duckweed and algal ponds, *Water Sci. Technol.*, *42*(10), 363–370.

Vidon, P., L. P. Tedesco, J. Wilson, M. A. Campbell, L. R. Casey, and M. Gray (2008), Direct and indirect hydrological controls on concentration and loading in midwestern streams, *J. Environ. Qual.*, *37*(5), 1761–1768, doi:10.2134/jeq2007.0311.

Walker, F. R., and J. R. Stedinger (1999), Fate and transport model of Cryptosporidium, *J. Environ. Eng.*, *125*, 325–333.

Whitman, R. L., M. B. Nevers, G. C. Korinek, and M. N. Byappanahalli (2004), Solar and temporal effects on *Escherichia coli* concentration at a Lake Michigan swimming beach, *Appl. Environ. Microbiol.*, *70*, 4276–4285.

Wilkinson, J., A. Jenkins, M. Wyer, and D. Kay (1995), Modelling faecal coliform dynamics in streams and rivers, *Water Res.*, *29*, 847–855.

Xie, Y., D. Lord, and Y. Zhang (2007), Predicting motor vehicle collisions using Bayesian neural network models: An empirical analysis, *Accid. Anal. Prev.*, *39*(5), 922–933.