WATER RESEARCH 42 (2008) 2563–2573

Available at www.sciencedirect.com

ScienceDirect

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

Water quality prediction of marine recreational beaches receiving watershed baseflow and stormwater runoff in southern California, USA

Li-Ming (Lee) He[a,*], *Zhen-Li He*[b]

[a]*California Environmental Protection Agency, Department of Pesticide Regulation, Surface Water Protection Program, 1001 I Street, P.O. Box 4015, Scramento, CA 95812, USA*

[b]*University of Florida, Institute of Food and Agricultural Sciences, Indian River Research and Education Center, 2199 South Rock Road, Fort Pierce, FL 34945, USA*

ARTICLE INFO

Article history:
Received 14 June 2007
Received in revised form
24 December 2007
Accepted 2 January 2008
Available online 5 January 2008

Keywords:
Artificial neural network
Beach water quality
Fecal indicator bacteria
Real-time prediction
Stormwater runoff

ABSTRACT

Beach advisories are issued to the public in California when the concentration of fecal indicator bacteria (FIB), including total coliform, fecal coliform (or *Escherichia coli*), and *Enterococcus*, exceed their recreational water health standards, or when the amount of a rainfall event is above the pre-determined threshold. However, it is not fully understood about how and to what degree stormwater runoff or baseflow exerts impacts on beach water quality. Furthermore, current laboratory methods used to determine the FIB levels take 18–96 h, which is too slow to keep pace with changes in FIB levels in water. Thus, a beach may not be posted when it is contaminated, and may be posted under advisory when bacterial levels have already decreased to within water quality standards. The study was designed to address the above critical issues. There were large temporal and spatial variations in FIB concentrations along two popular State Beaches in San Diego, CA, USA. The rainstorm-induced runoff from the watersheds exerts significant impacts on the marine recreational water quality of the beaches adjacent to lagoons during the first 24–48 h after a rain event. The large volume of stormwater runoff discharging to beaches caused high FIB concentrations in beach water not only at the lagoon outlet channel and the mixing zone, but also at the locations 90 m away from the channel northward or southward along the shoreline. The geomorphology of beach shoreline, distance from the outlet channel, wind strength, wind direction, tide height, wave height, rainfall, time lapse after a rainstorm, or channel flow rate played a role in affecting the distribution of FIB concentrations in beach water. Despite the great temporal and spatial variability of FIB concentrations along a shoreline, the artificial neural network-based models developed in this study are capable of successfully predicting FIB concentrations at different beaches, different locations, and different times under baseflow or rainstorm conditions. The models are based on readily measurable variables including temperature, conductivity, pH, turbidity, channel water flow, rainfall, and/or time lapse after a rainstorm. The established models will help fill the current gap between beach posting and actual water quality and make more meaningful and effective decisions on beach closures and advisories.

Published by Elsevier Ltd.

*Corresponding author. Tel.: +1 916 327 7479; fax: +1 916 324 4088.
E-mail addresses: lhe@cdpr.ca.gov (L.M. He), zhe@ufl.edu (Z.L. He).

0043-1354/$ - see front matter Published by Elsevier Ltd.
doi:10.1016/j.watres.2008.01.002

1. Introduction

Coastal beaches in California attract multimillions of visitors from all over the world. While much effort has been made during the past decade to protect and improve beach water quality, coastal water pollution remains a significant public health concern in southern California. It was estimated that between 627,800 and 1,479,200 excess gastrointestinal illnesses occur at beaches in Los Angeles and Orange Counties each year, corresponding to an annual economic loss of $21 or $51 million (Given et al., 2006) due to the illnesses. More than 5000 beach closing and health advisory days were reported across California in 2005. Beach advisories are generally issued to the public for 72 h after 0.2 or more inches of rain or when fecal indicator bacteria (FIB), including total coliform (TC), fecal coliform (FC) (or *Escherichia coli*), and *Enterococcus* (EN), exceed recreational water quality objectives established by the State of California. Although it is well known that stormwater runoff carries a significant level of FIB and may result in exceedance of recreational water quality objectives, it is not fully understood as to how and to what degree stormwater runoff from a watershed discharging to a coastal beach through a lagoon impacts recreational water quality temporally and spatially due to varying input rates, nearshore hydrodynamics, and bacterial die-off rates. There are six lagoon systems that directly discharge to recreational beaches in San Diego County. These systems drain large areas of agricultural, residential, commercial, and/or municipal lands that are potential bacterial sources. To better manage recreational water use, there is a need to better understand the spatial and temporal variability of beach water quality under the influence of baseflow and stormwater runoff discharging from watersheds.

TC, FC, or EN concentrations, on which beach advisories or closures are based, are currently determined by water sample collection and subsequent laboratory analysis. Laboratory procedures used to determine the FIB levels usually take 18–96 h, which is too slow to keep pace with changes in FIB levels in water (Hou et al., 2006; Kim and Grant, 2004). The monitoring data in San Diego County indicated that three quarters of contaminated beach waters were clean 24 h later. Thus, a beach remains open when it is contaminated, whereas the beach is closed when FIB levels have already decreased to within water quality standards. During this period of time, beach goers might have been exposed to harmful pathogens in water. This is the scenario of how recreational beach managers are currently operating. This operating system has been referred to as "persistence model", which assume that today's FIB levels are equivalent to yesterday's.

This issue could be resolved by a rapid instrumental method that can be used to measure indicator bacteria, but such a technique still requires 2–4 h and is at the early stage of implementation and validation. Even when these molecular techniques are commercially available for rapid measurement of FIB, they will not provide protection to swimmers prior to sampling and during sample processing due to rapid temporal variations of FIB concentrations in marine beach water (Boehm et al., 2002, 2007).

An alternative is to develop a water quality predictive model that can be used to predict the levels of FIB in beach water in near real time. Rainfall-based regression models have been used for pre-emptive beach closures in Milwaukee, Wisconsin; Stamford, Connecticut; Sussex, Delaware, and Boston, Massachusetts (Boehm et al., 2007; USEPA, 2002). Since the rainfall-based linear models assume linear relationships and are site dependent, they may not provide desirable results when applied to both baseflow and rainstorm conditions considering the variability of weather, land use, watershed characteristics, and coastal water hydrodynamics in southern California. The rainfall pattern in southern California is significantly affected by the El Nino Southern Oscillation that modulate rainfall and stormwater runoff (Pednekar et al., 2005). It is hypothesized that nonlinear approaches such as artificial neural networks (ANNs) are adequate for predicting beach bacterial concentrations in complex systems. Previous studies indicated that ANNs were successfully used to predict or forecast water quality parameters such as algal concentration and nutrient levels in freshwater and marine waters, and FIB source differentiation as they are able to capture nonlinear relationships among input and output variables (Brion et al., 2002; Kuo et al., 2006; Lee et al., 2003).

The objectives of this study were to characterize the temporal and spatial variations of FIB at coastal beaches adjacent to lagoons receiving discharges from watersheds under baseflow and rainstorms in southern California and to develop an ANN model for prediction of FIB in coastal water using readily *in-situ* measurable water quality parameters and other online accessible weather/hydrological variables. The approach developed in this paper may be used as a new tool to complement and improve current recreational water quality management practices in California, the US, and other countries over the world.

2. Materials and methods

2.1. Study sites

The study sites were located in San Diego County in southern California, USA (Fig. 1). The County encompasses 52 miles of beach recreational waters out of 154 total shoreline miles. Beach waters used for full body-contact recreational activities such as swimming, surfing and diving must meet specific bacteriological standards to be considered safe for such purposes. Many of the recreational beaches in San Diego receive direct discharges from stormdrains, creeks or rivers, or indirectly through lagoons. There are six lagoons adjacent to recreational beaches in the county, two of which were selected as study sites for this project: The Los Penasquitos Lagoon discharging to the Torrey Pines State Beach (TPSB), and the San Elijo Lagoon discharging to the San Elijo State Beach (SESB) (Fig. 1).

The two state beaches adjacent to lagoons were selected because of their urbanized land uses upstream within the watersheds and year-round recreational uses for approximately 1–2 million visitors yearly (Table 1). The TPSB is located at the northwestern border of the City of San Diego,

Fig. 1 – Location of two study sites—San Elijo and Torrey Pines State Beaches in San Diego County, CA, USA.

direct south of the City of Del Mar. The beach receives water from the Penasquitos Creek watershed, which encompasses 245 km^2. Major land uses within this watershed are residential (24.7%), undeveloped (17.8%), and parks and recreation (29.2%). Other uses are comprised of industrial (7.3%), commercial (2.5%), transportation (11.9%), and agriculture (1.6%) (Table 1).

The SESB is located 20 miles north of San Diego, between the cities of Solana Beach and Encinitas. The beach receives water from the Escondido Creek watershed, which covers ~220 km^2. Land uses within the watershed are predominantly undeveloped (35%), residential (25%), and parks (16%) (Table 1).

The residential land use in the Los Penasquitos Creek watershed is similar to that in the Escondido Creek watershed, whereas the former seems more urbanized that the latter that showed higher agricultural and undeveloped land uses. While comparing to the overall land uses in San Diego County, the two watersheds in the study exhibited higher residential and lower undeveloped land uses.

2.2. *Sampling*

A water quality monitoring plan for the project was developed for water sampling, field measurement, laboratory analysis, and quality control and assurance. Water samples were collected on dry and wet events at the two state beaches. The dry event samples refer to those collected at least 7 days after a rainstorm while the wet event samples refer to those collected within 0–5 days after a rainstorm. During wet sampling events, daily samples were collected up to 5 days in the same tide window (2 h during ebbing prior to a peak low tide) to minimize the potential influence of tide height. The same sampling pattern was followed throughout the study, and samples at a specific location were taken at the same time window to minimize the effect of sunlight and salinity on the die-off of bacteria. Most samples were collected between 11:30 and 13:30 at SESB and between 13:30 and 15:00 at TPSB. Two rainstorm events were sampled in this study. The first rainstorm event, which was on March 15, 2003 had about 50 mm of rainfall, and the second event on April 14, 2003 had 45 mm of rainfall. During dry sampling events, samples were collected using the same protocol for wet event sampling. The rainfall data were obtained from the San Diego County's Flood Control District. There were multiple rain gauge stations within a watershed. The average rainfall from each of the two storm events was used in this study, which was averaged from rainfall recorded by multiple rain gauges within the watershed.

All water samples were taken at eight locations along the shoreline at each study beach (Fig. 2). These locations are the lagoon outlet channel (OLCH), the lagoon water and ocean water mixing zone (M0), 23 (L01), 45 (L02), and 90 (L03) meters southward from the mixing zone; 23 (R01), 45 (R02), and 90 (R03) meters northward from the mixing zone.

Water samples were taken 10–15 cm below the water surface in knee-deep water (~0.5 m) at all the locations. All beach water samples were collected on an incoming wave in 125-mL sterile plastic bottles, placed on ice, and delivered to the laboratory within 6 h. Lagoon outlet channel samples were collected as close to the center of the channel as can be safely accessed. A sampling rod was used to collect water samples. An uncapped bottle was submerged under water, rotated to side (90°), and swept horizontally keeping the bottle at even level under the water surface. Samples were delivered to the San Diego County's Public Health Laboratory for measuring TC, FC, and EN.

At each site when a water sample was taken, the physicochemical properties of water were determined using a portable multiprobe system (Horiba U-10, Japan). Water parameters included pH, temperature, conductivity, turbidity, and dissolved oxygen. Salinity was calculated from conductivity. The lagoon outlet flow rate was obtained by measuring the water depth, channel width, and flow velocity. The flow velocity was measured using a flow probe (FP101 Global Flow Probe, Gold River, CA).

During water sampling, the number of birds, visible bird waste, and the number of swimmers and surfers, if any, were recorded at the sampling site in the field. The direction of littoral current was estimated using the "orange method" (Boehm et al., 2003) by observing the direction of floating orange movement. Tide height was calculated at the time the sample was collected using WXTide32-a Windows tide and current prediction program (http://www.wxtide32.com/

Table 1 – Summary information of the two coastal beaches in southern California, USA

State Beach	Watershed Area (km^2)	Lagoon area (km^2)	Number of beach visitors (1989)	Land use (%)						
				Agricultural	Commercial	Industrial	Residential	Transportation	Parks	Undeveloped	Other
TPSB	245	1.72	1,024,000	1.6	2.5	7.3	24.7	11.9	29.2	17.8	5.0
SESB	219	1.68	1,809,000	8.9[a]	2.3[a]	3.1[a]	25.0	5.5[a]	16.0	35.0	6.6

[a] Estimated from the Carlsbad Watershed Management Area that included the Escondido Creek.

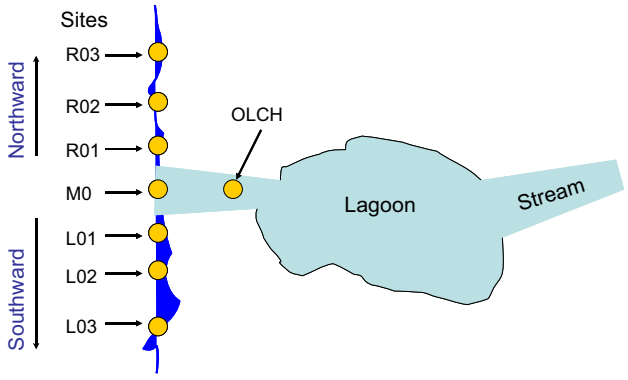

Fig. 2 – Schematic diagram for sampling locations (orange color dots).

index.html). The wave height was obtained from the Coastal Data Information Program (CDIP) (http://cdip.ucsd.edu).

The multiple tube fermentation (MTF) method (Standard Method 9221 B and E) was used for measuring TC and FC. Results were expressed as Most Probable Number (MPN) per 100 mL of water (MPN/100 mL). The IDEXX Enterolert method (Standard Method 9223 B Enzyme Substrate Test) was used for measuring EN. Results were expressed as MPN/100 mL. Enterolert uses a Defined Substrate Technology® (DST®) nutrient indicator to detect EN. This nutrient-indicator fluoresces when metabolized by EN. The 97-well Quanti-Tray was used for all samples or otherwise indicated.

2.3. *ANNs*

ANNs were applied to develop a model for predicting recreational water quality using physicochemical properties of water and weather variables readily available from monitoring. Since the ANN approach will not assume any functional relationship between the dependent and independent variables, ANNs are suitable for capturing functional relationships between bacterial levels and other variables in beach water.

The ANN used in this study was a fully connected, feed-forward system trained with a backpropagation algorithm (He et al., 2003). The software was NeuralWorks Predict or Professional II Plus (NeuralWare, Inc., Pittsburgh, PA). The backpropagation network uses supervised training (learning), which requires both the inputs and the outputs. The network processes the inputs and compares its resulting outputs against the desired outputs. Errors are then propagated back through the system, causing the system to adjust the weights (parameters) in each layer, which control the network. This process occurs over and over as the weights are continually tweaked. The set of data that enables the training is called the "training set." During the training of a network the same set of data is processed many times as the connection weights are ever refined. In the course of learning, the residual between the model output and the desired output decreases and the model learns the relation between the input and the output. Learning rules are written so that the iterative learning process minimizes the error measure.

The training must be stopped at the right time. If training continues for too long, it results in overlearning. Overlearning means that the neural network extracts too much information from the individual cases forgetting the relevant information of the general case. When the network starts to learn the characteristics on individual samples rather than the characteristics of the general phenomenon, the model residual for the testing set starts to increase, while the model residual still decreases for the training set. The model is departing from the general structure of the problem to learning about the individual cases instead. That is why the neural network performance is tested with a "testing set" that is not part of the training set. The testing set can be seen as the representative cases of the general phenomenon. If the network performs well on the testing set, it may be expected to perform well on the general case as well. When this step of traing and testing an ANN model is complete, the model is ready for next step—validation.

The ANN is then said complete and ready for validation. The data sets unused in both training and test are used to examine the fitness of the established ANN model.

2.4. *Data analysis and model validation*

During the processes of data analysis, model development and validation, FIB concentrations and other water quality parameters were log transformed where necessary. Non-detects (i.e., below a method detection limit (MDL)) were substituted with one half the MDL. All ANN models were developed by using the combination of the data collected from the two state beaches. The trained and tested ANN models were validated with a validation dataset that had not been used in the training and testing stages. The validation dataset for each developed ANN model consisted of data collected from the two state beaches.

For ANN model validation, predicted values from a validation dataset using the developed ANN model are plotted against the measured values to provide a graphical view of model fitness. Model fitness was also examined by using the following methods: Correlation coefficient, prediction interval, root mean square error, and false positive/negative rates. The correlation coefficient r or r^2 a measure of the strength of the relationship between the predicted and measured values.

The prediction intervals presents an interval estimate of a variabl and predict the distribution of individual points. The predicted value from an ANN model will have a margin of error. The predicted value is a point estimate and a confidence interval may be calculated for that estimate. This confidence interval is referred to as prediction interval. For example, the 90% prediction interval is the area in which 90% of all data points are expected to fall. While a predicted FIB concentration is plotted against a measured FIB concentration (e.g., Fig. 5), a perfect prediction would generate a perfect 1:1 line. For illustration and practical purpose, the 90% prediction interval bounds (lines) were used to show the fitness of an ANN model. If all data points fall within the 90% prediction interval bounds, the rate of correct prediction is 100%.

The root mean squared error (RMSE) is the square root of the mean square error. That is probably the most easily interpreted statistic, since it has the same units as FIB. The RMSE is thus the distance, on average, of a data point from the fitted line, measured along a vertical line.

From the regulatory perspective, the false positive/negative rate may be more useful. A false positive refers to a predicted FIB concentration that is higher than a water quality objective while the measured concentration is below the water quality objective. A false negative refers to a predicted FIB concentration that is lower than a water quality objective while the measured concentration is higher than the water quality objective.

3. Results and discussion

3.1. *Water quality measured in situ using a multiparameter meter*

On average, water temperature was higher at the outlet channel than at any other sites ($p<0.05$) (data not shown) while it was slightly lower at SESB than TPSB (Table 2). Conductivity and pH were lower at the outlet channel than at any other sites ($p<0.05$), whereas they were higher at SESB than at TPSB ($p<0.01$). Dissolved oxygen was not significantly different between the two beaches. Turbidity was higher at the outlet channel than at R02 and R03 ($p<0.05$). Turbidity is the indirect measurement of particulates suspended in water, which can originate from stormwater runoff and/or marine sediment resuspension by near shore tides. Turbidity was particularly high 1 day after a rainstorm at both state beaches.

At SESB, higher turbidity and lower conductivity and pH were observed on the south side than the north side. SESB exhibited a trough on the south side, which would allow more

Table 2 – Water quality properties at the two coastal beaches in southern California, USA

State beach	Statistic	Air temp (°C)	Water temp (°C)	pH	Conductivity (mS/cm)	Turbidity (NTU)	Dissolved oxygen (mg/L)	Log (TC)	Log (FC)	Log (EN)
TPSB	Mean	15.3	16.9	8.10	30.97	26.4	5.46	3.188	2.453	2.167
	StDev	1.1	0.9	0.14	15.71	35.4	0.51	0.971	0.849	1.008
TPSB	Mean	15.6	16.5	8.18	44.63	16.4	5.38	2.801	2.136	1.838
	StDev	1.3	1.0	0.15	9.54	11.8	0.59	0.894	0.773	0.785

water flowing from the lagoon to the trough on the south side, leading to higher turbidity and lower conductivity and pH on the south side. This geomorphology also affected bacterial distributions at SESB as discussed in the bacterial result section. The effect of wind and current direction on lagoon water plume distribution along the beach was observed for conductivity and pH. When the littoral current moved from north to south on March 17 and April 15, conductivity and pH were higher on the north side than the south side. On the contrary, on April 17, the southerly current resulted in lower conductivity and pH values on the north side.

3.2. *FIB at two state beaches*

In general, the outlet channel and the mixing zone showed high FIB concentrations at both state beaches and their levels decreased with an increase in distance from the channel (Fig. 3). FIB concentrations were higher at TPSB than at SESB ($p<0.05$) (Table 2) and the pattern of FIB distribution along the shoreline was different between the two beaches (Fig. 3). FIB levels at TPSB were similar on north and south sides as TPSB exhibited a relatively straight and level shoreline and the water from the lagoon would be expected to be evenly distributed to both south and north sides through rip currents. The TC concentration at TPSB decreased as the distance from the outlet channel increased (Fig. 3), indicating dilution and/or die-off of TC during the along-shore transport process. However, the magnitude of changes in both FC and EN with distance decreased. The average EN levels at each site at TPSB appeared similar at all sites except for the site R03. This may indicate the better survival rate of EN in marine water.

The FIB levels at SESB were higher on the south side than the north side (Figs. 3 and 4b). The difference may be caused by various factors including the geomorphology of the beach. SESB had a trough on the south side and more water was expected to flow from the lagoon to the trough on the south side, leading to higher bacterial concentrations as well as lower conductivity and higher turbidity on the south side. Furthermore, FIB concentrations were particularly high at the site L03 that was 90 m away from the outlet channel. This could result from the combination of geomorphology and fine

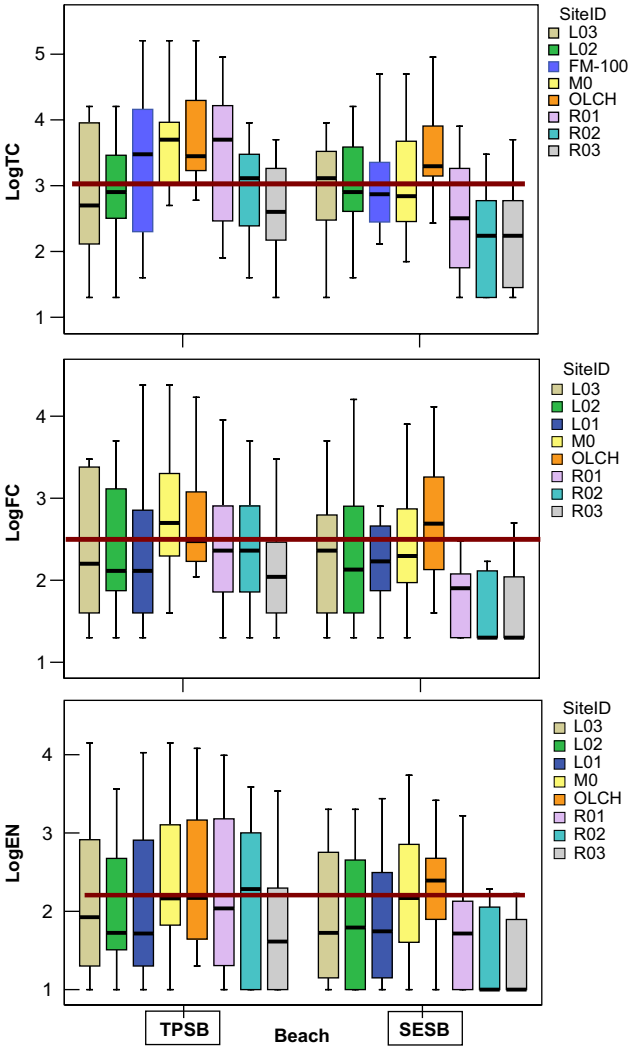

Fig. 3 – Variability of total coliform (TC), fecal coliform (FC), and *Enterococcus* (EN) at different locations on two State beaches in southern California. The solid line indicates recreational water health standards. Box—25th and 75th percentiles. Whisker—1 and 99th percentiles. The line in the box—median.

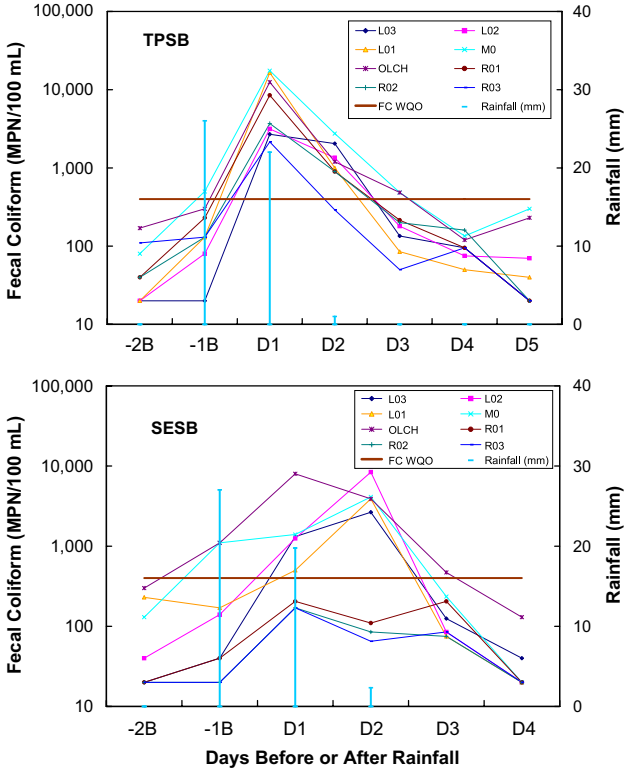

Fig. 4 – Changes in fecal coli form levels with time after a storm event at TPSB and SESB. The solid line indicates recreational water quality objective for fecal coliform. –2B—2 days before the rainstorm. D2—2 days after the rainstorm.

sediment at the slightly depressed site that would harbor excessive FIB. Indeed, the fine sediment at L03 resulted in high turbidity measurements at the site.

Correlation analysis (the Pearson correlation) indicated that TC, FC, and EN were significantly correlated to air temperature ($-0.41<r<-0.30$, $p<0.01$), pH ($-0.49<r<-0.43$, $p<0.01$), conductivity ($-0.73<r<-0.70$, $p<0.01$), turbidity ($0.61<r<0.66$, $p<0.01$), tide height ($-0.25<r<-0.17$, $p<0.05$), wave height ($0.44<r<0.52$, $p<0.01$), wind ($0.30<r<0.37$, $p<0.01$), wind direction ($0.29<r<0.36$, $p<0.01$), distance to the outlet channel ($-0.37<r<-0.22$, $p<0.01$), channel flow rate ($0.34<r<0.43$, $p<0.01$), time lapse from last rain ($-0.58<r<-0.44$, $p<0.01$), and rainfall amount ($0.42<r<0.58$, $p<0.01$). However, they were not correlated with the number of birds observed at beach sites ($-0.036<r<-0.002$, $p>0.6$). A recent study indicated that the number of FIB exceedances at Huntington and Newport Beaches in southern California was correlated with salinity but not with temperature, chlorophyll fluorescence, or tidal range. However, the number of exceedances was correlated with the Fisher Information and Shannon Entropy calculated from the measurements of salinity and temperature in the surf zone (Jeong et al., 2006).

FIB concentrations were greatly affected by rainstorm-induced runoff. The concentration reached highest one day after a rainstorm and then gradually decreased to baseline levels in 4 or 5 days (Fig. 4). Rainstorms generate urban runoff that wash and carry various constituents including FIB. FIB concentrations decreased as runoff flow decreased with time after a rainstorm due largely to the reduced flow to the beaches and potential bacterial die-off in marine water. The large amount of runoff water can be readily transported along the shoreline by strong rip currents and by northerly or southerly waves, resulting in high bacterial concentrations even beyond 90 m from the outlet channel after a rainstorm (Fig. 4).

There are strong correlations among TC, FC, and EN at the two study beaches. The correlations at TPSB are similar to those at SESB. On the average, the concentration of FC can be estimated from the TC concentration by the factor of 0.76 ($R^2=0.737$, $p<0.01$). The EN concentration can be estimated from the concentrations of TC or FC by the factors of 0.68 ($R^2=0.705$, $p<0.01$) or 0.89 ($R^2=0.781$, $p<0.01$), respectively. This suggests that the level of one indicator may be reliably estimated from the result of another indicator. While comparing with the recreational water quality criteria established by the California Department of Health Services: 10,000 (1000 if TC:FC<10) MPN/100 mL for TC, 400 MPN/100 mL for FC, and 104 MPN/100 mL for EN, the highest number of bacterial exceedances at the two State beaches is EN, followed by FC. The average TC:FC ratio was 10.97 at TPSB and 8.54 at SESB, and there was no significant difference between the two beaches using *t*-test ($p=0.313$). All the three indicator bacteria were below the marine recreational water quality criteria 3 days after a rainstorm (Fig. 4).

3.3. *FIB prediction using ANNs*

ANNs were employed to predict FIB concentrations using readily measurable parameters. Our first attempt was to establish an ANN model for simultaneous prediction of TC, FC, and EN. The second attempt was to construct separate ANN models for the prediction of individual TC, FC, or EN. The ANN constructed to predict all three FIB simultaneously had five input variables including water temperature, conductivity, turbidity, time lapse from last rain, and rainfall (Table 3). The architecture of the ANN (ANN-1) consisted of seven inputs, six hidden-layer neurons, and three outputs (TC, FC, and EN) (Table 3). At the 90% prediction interval, indicator bacteria were correctly predicted with the training dataset at a rate of 94% for TC, 90% for FC, and 91% for EN (Table 3). With the test dataset, concentrations were correctly predicted at 100% for TC, 93% for FC, and 97% for EN. With the validation dataset, concentrations were correctly predicted at 97% for TC, 94% for FC, and 89% for EN (Table 4 and Fig. 5).

Model validation is possibly the most important step in the model building sequence. It is also one of the most overlooked. There are many statistical tools for model validation, but the primary tools for most process modeling applications include correlation coefficient (r^2), prediction intervals, RMSE, and false positive or false negative rates. From the perspective of modeling accuracy, r^2, prediction intervals, and RMSE are better indicators for assessing model fitness. From the

Table 3 – Input, output, and architecture of artificial neural networks developed for predicting fecal indicator bacteria levels in two coastal beaches in southern California, USA

ANN	Input variables selected	Output	Architecture[a]	Overall R[b]
ANN-1	Water temperature, conductivity, turbidity, last rain, rainfall	TC, FC, EN	7-6-3	0.845 (TC), 0.859 (FC), 0.909 (EN)
ANN-2	Water temperature, pH, conductivity, tide height, wave height, last rain	TC	7-3-1	0.893
ANN-3	Water T, pH, conductivity, flow rate, last rain, rainfall	FC	12-6-1	0.907
ANN-4	pH, conductivity, tide height, wave height, last rain, rainfall	EN	7-8-1	0.932

[a] The architecture of an ANN model is the ANN structure that presents the number of neurons in each of three layers—input, hidden, and output.
[b] R is the linear correlation between predicted values and measured values. The overall R is the average of three R's for training, test, and validation, respectively. All FIB data were log transformed.

Table 4 – Summary of artificial neural networks developed for predicting fecal indicator bacteria levels in two coastal beaches in southern California, USA

ANN	Dataset	R^2	RMSE	Total number of samples	Number of samples within 90% CI	Percent samples within 90% CI
ANN1-TC	Train	0.628	0.504	118	111	94.1
	Test	0.715	0.456	30	30	100.0
	Validation	0.620	0.553	36	35	97.2
ANN1-FC	Train	0.699	0.432	118	106	89.8
	Test	0.694	0.417	30	28	93.3
	Validation	0.762	0.424	36	34	94.4
ANN1-EN	Train	0.855	0.353	118	107	90.7
	Test	0.726	0.476	30	29	96.7
	Validation	0.821	0.454	36	32	88.9
ANN2-TC	Train	0.761	0.408	103	96	93.2
	Test	0.768	0.430	45	42	93.3
	Validation	0.731	0.475	36	34	94.4
ANN3-FC	Train	0.795	0.341	103	98	95.1
	Test	0.728	0.395	45	42	93.3
	Validation	0.833	0.307	36	36	100.0
ANN4-EN	Train	0.883	0.306	103	105	92.2
	Test	0.878	0.346	45	38	84.4
	Validation	0.789	0.407	36	34	94.4

regulatory perspective, the false positive and false negative rates will better reflect the usefulness of a model, indicating the uncertainty of the model used for compliance.

The developed ANN model was validated using a validation dataset. The validation dataset, consisting of 36 samples taken from the two state beaches, was randomly selected from the entire dataset to include samples representative of the entire dataset such as data range, weather condition, sample location, and channel water flow in these two State beaches. It is worth noting that the validation dataset included samples taken from the beginning to end, with a number of these samples collected during the final days of sampling for the project. Results from the model validation indicate that the developed ANN model is capable of correctly predicting FIB concentrations at or above 90% (Table 4). Their correlation coefficients and RMSEs are similar to those from training or test datasets. Out of 36 validation samples, there were two false positives and two false negatives for TC, two false positives (individual samples) and one false negative for FC, and no false positives and two false negatives for EN (Fig. 5). While ANNs were constructed to predict TC, FC, or EN individually, the ANN for TC (ANN-2) had six input variables including water temperature, pH, conductivity, tide height, wave height, and time lapse from last rain (Table 2). The architecture of the ANN consisted of seven inputs, three hidden-layer neurons, and one output (TC). At the 90% prediction interval, TC was correctly predicted at 93% with the training dataset, at 93% with the test dataset, and at 94% with the validation dataset (Table 4). The ANN for FC (ANN-3) had six input variables including water temperature, pH, conductivity, flow rate, time lapse from last rain, and rainfall amount (Table 4). The architecture of the ANN consisted of 12 inputs, six hidden layer neurons, and one output (FC). At the 90% prediction interval, FC was correctly predicted at 95% with the training dataset, at 93% with the test dataset, and at 100% with the validation dataset (Table 4). The ANN for *EN* (ANN-4) had six input variables including pH, conductivity, tide height, wave height, and time elapse from last rain, and rainfall amount (Table 4). The architecture of the ANN consisted of seven inputs, eight hidden-layer neurons, and one output (EN). At the 90% confidence interval, EN was correctly predicted at 91% with the training dataset, at 84% with the test dataset, and at 94% with the validation dataset (Table 4). As the ANNs for individual TC, FC, and EN prediction failed to provide better prediction of TC, FC, or EN at the 90% prediction interval, the ANN model for simultaneous prediction of TC, FC, and EN is preferred due to its efficacy and simplicity.

A close look at explanatory variables selected as the inputs to ANN models can provide insights into the relative importance of these variables. Input variable selection from a total of 16 explanatory variables was completed in two steps: cascaded and genetic variable selection. The cascaded variable selection uses probabilities to reduce the size of variables, eliminating those variables that have a very low probability of inclusion in an optimum solution. Immediately followed is the second variable selection technique—genetic variable selection. This technique uses a logistic multiple regression algorithm to select the model's input variables. Two variables, conductivity and time lapse from last rain, were selected by all four ANN models independently developed in this study (Table 3). Three variables—water temperature, pH, and rainfall were used in three models, whereas other three variables—channel flow rate, tide height, and

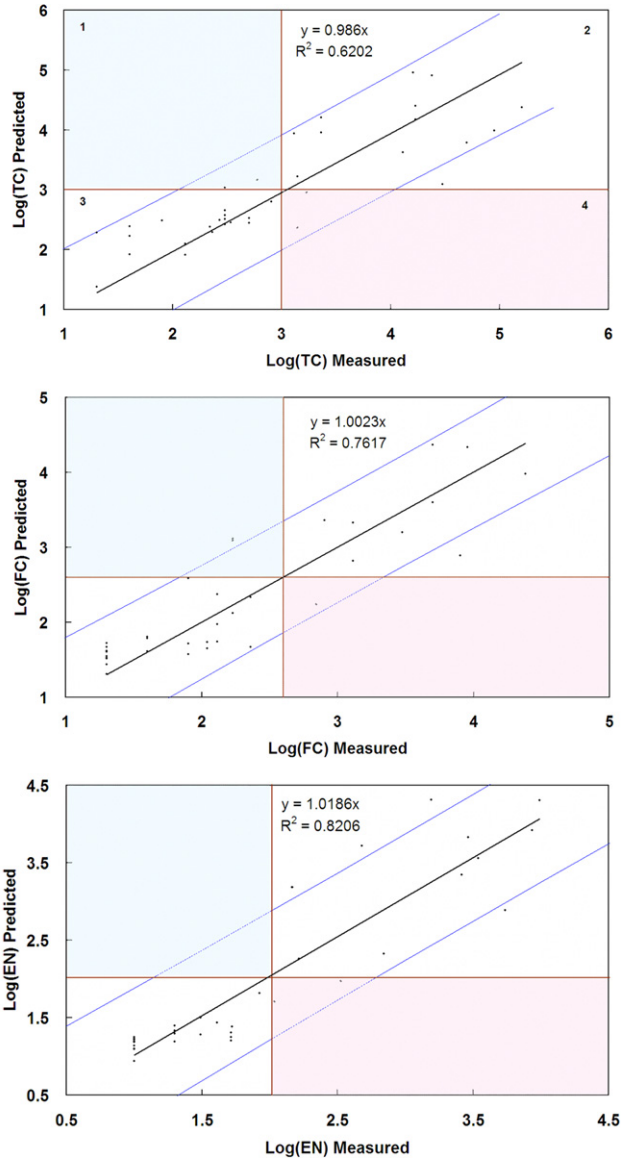

Fig. 5 – Validation of the artificial neural network model developed for simultaneous prediction of total coliform (TC), fecal coliform (FC), and *Enterococcus* (EN) and 90% confidence intervals in two coastal beaches in southern California, USA.

wave height were employed by two models. And turbidity was selected by one model. It has been substantiated that FIB die-off rates are higher in water with higher conductivity levels (Boehm, 2003; Easton et al., 2004; Oliver et al., 2006), which in turn are affected by channel flow rates regulated by rainfall and time lapse from last rain event. Rainstorms generate surface runoff when they exceed a threshold of water holding capacity of the surface. Runoff water carries a variety of contaminants including FIB from residential, agricultural, or industrial land-use areas, which contribute to elevated contaminant levels in receiving water. Surface water temperature is most likely affected by daily solar radiation—higher in sunny days, which adversely affect FIB concentrations due largely to higher UV radiations. Tide or wave height exerts impacts on the movement of near-shoreline sediment, which may be the place that harbors a significant amount of FIB (Boehm et al., 2007; Gruber et al., 2005; He et al., 2007; Ishii et al., 2007; Lee et al., 2006; Yamahara et al., 2007). Turbidity was selected into the model that simultaneously predicts all three FIB. However, turbidity was not selected by other three models developed to predict TC, FC, or EN separately. The exclusion of turbidity from three individual models does not imply its lack of importance, but may indicates less importance than other parameters, which were selected based on certain threshold values used in the variable selection process. Furthermore, since turbidity is very much impacted by stormwater runoff and/or tide height and wave height during the wet season, the inclusion of tide height, wave height, rainfall, or time elapse from last rainfall in a model would have carried the information the turbidity data may have.

Other nonlinear models have been developed to predict bacterial levels in beach water (Boehm et al., 2007; Parkhurst et al., 2005). Parkhurst et al. (2005) applied a nonlinear approach to predict the concentrations of *E. coli* and EN in five recreational beaches including the Imperial Beach in San Diego, California. They used the random forests method, which is an extension of tree regression, to explore the relationships between FIB levels and other explanatory variables including information on weather, tide, wave, day of week, time of day, sampling depth, and FIB concentrations 24 h earlier. The results from their study indicated that important predictors differed substantially among the five beaches, but day of week, concentrations 24 h earlier, and sampling depth were strongly related to the levels of *E. coli* and EN. For FIB levels, the approach worked poorly for raw data of bacterial concentrations, whereas prediction was improved considerably when the log-transformed data were used, with an R^2 as high as 0.82.

While comparing modeling results from various linear or nonlinear regression techniques used in the Huntington State Beach study, Boehm et al. (2007) found that partial least square (PLS) regression models performed better than others. The regression trees provided better prediction with the training dataset than the validation dataset for EN. The regression trees performed better in wet seasons than dry seasons based on a comparison of error rates to the persistence model (Boehm et al., 2007; Yamahara et al., 2007).

The ANN-based approach reported here provides another simple and reliable way to predict beach water quality using input variables (e.g., pH, conductivity, turbidity, etc.), which were not used in the Parkhurst's approach, can be readily measured *in situ* or available online in real time. The approach could be employed to provide instant warnings of beach water contamination while used with discrete water quality monitoring such as a multiparameter water quality meter as described in this paper. Previously, simple linear regression or PLS regression models were applied to estimate bacterial levels in rivers and coastal beaches. Their findings indicated that bacterial levels could be estimated with physicochemical properties and weather conditions at Ohio bathing beaches (Francy et al., 2003). The PLS model produced a more favorable balance between illness prevention and recreational access at coastal beaches in southern California (Hou et al., 2006). The

ANN models presented in this paper provides predictive results with few false positives or false negatives (<10%).

The approach may also be used to predict or forecast bacterial levels in beach water in near real time when coupled with continuous water quality monitoring systems. The continuous monitoring systems have been developed and deployed in streams or coastal waters for near-real-time measurement of the physicochemical properties of water. The readily measurable and continuous water quality data (e.g., temperature, pH, conductivity, dissolved oxygen, turbidity, or chlorophyll) may be used as surrogates to estimate other difficult-to-measure constituents (e.g., bacteria or nutrients) as demonstrated with simple regression models for watersheds in Kansas (Rasmussen et al., 2005). The coastal ocean observing systems deployed in southern California provide monitoring data with high temporal resolution and spatial coverage, and easy accessibility (Jeong et al., 2006) and represent opportunities for developing and implementing a predictive or forecasting model system for water quality management in near real time. Even though the predictive models are developed to predict indicator bacteria in water, they may be extended to assess potential human health risks in terms of water-borne pathogens as strong correlations exist between FIB and pathogenic viruses (i.e., hepatitis a virus and enteroviruses) in water polluted with a combination of possible sewage and stormwater runoff (Gersberg et al., 2006). As these ANN models are integrated into the deployed real-time monitoring systems, the approach described here may therefore be used as a new tool to complement and improve current recreational water health management practices in California and the nation alike.

While the ANN models were developed using water quality data collected during the wet season, models developed for dry weather seasons may be more important as more people go to beaches and beach water quality may change more frequently and randomly, becoming more difficult to predict for the summer dry season than for the wet season. Even though some of the data used in the ANN model development were collected under baseflow conditions, it is anticipated that beach water quality in the dry season differs from that in the wet season since pollution sources in the dry season are likely more diverse. It will be useful to develop a beach water quality prediction model for summer seasons. This requires water sampling and data collection during the dry season.

The principle used in the development and validation of predictive models may be extended to other aspects of coastal water protection. Many coastal waterbodies serve for coastal recreation and tourism, whereas many others provide a unique, highly productive environment that supports a great diversity of wildlife and fisheries and contributes tremendous value to the economy. Coastal habitats provide spawning grounds, nurseries, shelter and food for finfish, shellfish, birds and other wildlife, as well as nesting, resting, feeding, and breeding habitat for 85% of waterfowl and other migratory birds in the US. Estuaries provide habitat for more than 75% of America's commercial fish catch, and for 80–90% of the recreational fish catch (USEPA, 2007). Similar ANN models can be developed to predict or forecast pathogen levels in shellfish beds to protect from consuming fecal contaminated shellfish.

This modeling technique can also be used to develop models for predicting or forecasting the potential for coastal harmful algal blooms. The toxins released from harmful algae will kill wild and farmed fish and shellfish, cause human illness and death from contaminated shellfish or fish, and result in death of marine mammals, seabirds, and other animals (Anderson, 1995). The best way to control the spread of harmful algae is to stop or slow their growth at the early stage of their occurrence. Harmful algal blooms are likely controllable while they are contained in a small embayment. If we can predict or forecast a harmful algal bloom, control measures will be more successful. The ANN models may be used as a tool to reduce the damage caused by harmful algal blooms. In fact, ANNs have been used to develop predictive models for harmful algal blooms in lakes, rivers and coastal waters (Lee et al., 2003; Recknagel, 1997; Recknagel et al., 1997; Walter et al., 2001; Wei et al., 2001; Whitehead et al., 1997; Yabunaka et al., 1997).

4. Conclusions

Recreational beaches are subject to closure or posting advisory warning signs when the amount of rainfall exceeds a threshold or fecal indicator bacterial levels in beach water exceed established water quality criteria. Considering that it takes at least 18 h to obtain bacterial analytical results from a laboratory, beach water quality may have already changed so that a beach is posted when bacterial levels are below water quality criteria. In this study, we have developed a reliable ANN model to predict three FIB (TC, FC, and EN) simultaneously in two popular State beaches in San Diego, California. Explanatory variables selected to establish this model include water temperature, conductivity, turbidity, rainfall, and time lapse from last rainfall. These parameters can be readily measured *in situ* or in real time through online access to these data. The results from the model validation process show that the developed model makes accurate simultaneous prediction of the levels of three FIB, with false positive or negative rates less than 10%. The established model will help make rapid and effective decisions about beach closures or advisories and fill the current gap between beach posting and water quality conditions.

Although this model was developed using beach water quality data from two San Diego beaches, the wide temporal and spatial variations in bacterial concentration, stream flow, rainfall, and other water quality properties used in model development may suggest that the model is readily applicable to other coastal beaches with similar conditions. There is a potential that these ANN-based models can be applied to recreational waters with different conditions since ANNs have the capability of robust and generalized prediction. These models will be improved for application to multiple beaches having different conditions as more data collected from other recreational beaches are used for model development and validation.

Acknowledgment

We gratefully acknowledge those involved in monitoring design and data collection for this study, with special thanks

to Clay Clifton, Chattral Koether, and Jo Ann Weber. This project was funded, in part, by the US Environmental Protection Agency (EPA) under the Beaches Environmental Assessment and Coastal Health (BEACH) Act. The views expressed by the authors of the paper are their own and do not necessarily reflect the views and policies of US EPA. Mention of trade names, products, or services does not constitute endorsement or recommendation for use.

REFERENCES

Anderson, D.M., 1995. Toxic red tides and harmful algal blooms: a practical challenge in coastal oceanography. Rev. Geophysics, Suppl. US National Report to the International Union of Geodesy and Geophysics, 1991–1994:1189-1200.

Boehm, A.B., 2003. Model of microbial transport and inactivation in the surf zone and application to field measurements of total coliform in Northern Orange County, California. Environ. Sci. Technol. 37, 5511–5517.

Boehm, A.B., Grant, S.B., Kim, J.H., Mowbray, S.L., McGee, C.D., Clark, C.D., Foley, D.M., Wellman, D.E., 2002. Decadal and shorter period variability of surf zone water quality at Huntington Beach, California. Environ. Sci. Technol. 36, 3885–3892.

Boehm, A.B., Fuhrman, J.A., Mre, R.D., Grant, S.B., 2003. Tiered approach for identification of a human fecal pollution source at a Recreational Beach: case study at Avalon Bay, Catalina Island, California. Environ. Sci. Technol. 37, 673–680.

Boehm, A.B., Whitman, R.L., Nevers, M.B., Hou, D., Weisberg, S.B., 2007. Now-casting recreational water quality. In: Wymer, L., Dufour, A. (Eds.), Statistical Framework for Water Quality Criteria and Monitoring.

Brion, G.M., Neelakantan, T.R., Lingireddy, S., 2002. A neural-network-based classification scheme for sorting sources and ages of fecal contamination in water. Water Res. 36, 3765–3774.

Easton, J.H., Lalor, M., Gauthier, J.J., Pitt, R., 2004. In-situ die-off of indicator bacteria and pathogens. In: National Beach Conference, October 13–15, 2004, San Diego, California.

Francy, D.S., Gifford, A.M., Darner, R.A., 2003. *Escherichia coli* at Ohio Bathing Beaches—distribution, sources, wastewater indicators, and predictive modeling. US Geological Survey Water-Resources Investigations Report 02-4285, 117pp.

Gersberg, R.M., Rose, M.A., Robles-Sikisaka, R., Dhar, A.K., 2006. Quantitative detection of Hepatitis A virus and Enteroviruses near the United States-Mexico Border and Correlation with levels of fecal indicator bacteria. Appl. Environ. Microbiol. 72, 7438–7444.

Given, S., Pendleton, L.H., Boehm, A.B., 2006. Regional public health cost estimates of contaminated coastal waters: a case study of Gastroenteritis at Southern California Beaches. Environ. Sci. Technol. 40, 4851–4858.

Gruber, S., Aumand, L., Martin, A., 2005. Sediments as a reservoir of indicator bacteria in a coastal embayment: Mission Bay, California. Technical Paper #0506, 7pp., Weston Solutions, Carlsbad, California.

He, L.M., Kear-Padilla, L.L., Lieberman, S.H., Andrews, J.M., 2003. Rapid in situ determination of total oil concentration in water using ultraviolet fluorescence and light scattering coupled with artificial neural networks. Anal. Chim. Acta 478, 245–258.

He, L.-M., Lu, J., Shi, W., 2007. Variability of fecal indicator bacteria in flowing and ponded waters in Southern California: implications for bacterial TMDL development and implementation. Water Res. 41 (14), 3132–3140.

Hou, D., Rabinovici, S.J.M., Boehm, A.B., 2006. Enterococci predictions from partial least squares regression models in conjunction with a single-sample standard improve the efficacy of beach management advisories. Environ. Sci. Technol. 40, 1737–1743.

Ishii, S., Hansen, D.L., Hicks, R.E., Sadowsky, M.J., 2007. Beach sand and sediments are temporal sinks and sources of *Escherichia coli* in Lake Superior. Environ. Sci. Technol. 41, 2203–2209.

Jeong, Y., Sanders, B.F., Grant, S.B., 2006. The information content of high-frequency environmental monitoring data signals pollution events in the Coastal Ocean. Environ. Sci. Technol. 40, 6215–6220.

Kim, J.H., Grant, S.B., 2004. Public mis-notification of coastal water quality: a probabilistic evaluation of posting errors at Huntington Beach, California. Environ. Sci. Technol. 38, 2497–2504.

Kuo, J.-T., Hsieh, M.-H., Lung, W.-S., She, N., 2006. Using artificial neural network for reservoir eutrophication prediction. Ecol. Model. 200, 171–177.

Lee, J.H.W., Huang, Y., Dickman, M., Jayawardena, A.W., 2003. Neural network modelling of coastal algal blooms. Ecol. Model. 159, 179–201.

Lee, C.M., Lin, T.Y., Lin, C.C., Kohbodi, G.A., Bhatt, A., Lee, R., Jay, J.A., 2006. Persistence of fecal indicator bacteria in Santa Monica Bay beach sediments. Water Res. 40, 2593–2602.

Oliver, D.M., Haygarth, P.M., Clegg, C.D., Heathwaite, A.L., 2006. Differential E. coli die-off patterns associated with agricultural matrices. Environ. Sci. Technol. 40, 5710–5716.

Parkhurst, D.F., Brenner, K.P., Dufour, A.P., Wymer, L.J., 2005. Indicator bacteria at five swimming beaches—analysis using random forests. Water Res. 39, 1354–1360.

Pednekar, A.M., Grant, S.B., Jeong, Y., Poon, Y., Oancea, C., 2005. Influence of climate change, tidal mixing, and watershed urbanization on historical water quality in Newport Bay, a Saltwater Wetland and Tidal Embayment in Southern California. Environ. Sci. Technol. 39, 9071–9082.

Rasmussen, T.J., Ziegler, A.C., Rasmussen, P.P., 2005. Estimation of constituent concentrations, densities, loads, and yields in Lower Kansas River, Northeast Kansas, Using Regression Models and Continuous Water-Quality Monitoring, January 2000 Through December 2003. US Geological Survey Scientific Investigations Report 2005-5165, 117pp.

Recknagel, F., 1997. ANNA—artificial neural network model for predicting species abundance and succession of blue-green algae. Hydrobiologia 349, 47–57.

Recknagel, F., French, M., Harkonen, P., Yabunaka, K.I., 1997. Artificial neural network approach for modeling and prediction of algal blooms. Ecol. Model. 96, 11–28.

USEPA, 2002. Implementation Guidance for Ambient Water Quality Criteria for Bacteria EPA-823-F-02-009. Environmental Protection Agency, Washington, DC.

USEPA, 2007. National Estuary Program Condition Report, July 2007, EPA-842/B-06/001, 446pp.

Walter, M., Recknagel, F., Carpenter, C., Bormans, M., 2001. Predicting eutrophication effects in the Burrinjuck Reservoir (Australia) by means of the deterministic model SALMO and the recurrent neural network model ANNA. Ecol. Model. 146, 97–113.

Wei, B., Sugiura, N., Maekawa, T., 2001. Use of artificial neural network in the prediction of algal blooms. Water Res. 35, 2022–2028.

Whitehead, P.G., Howard, A., Arulmani, C., 1997. Modeling algal growth and transport in rivers—a comparison of time series analysis, dynamic mass balance and neural network techniques. Hydrobiologia 349, 39–46.

Yabunaka, K., Hosomi, M., Murakami, A., 1997. Novel application of a back-propagation artificial neural network model formulated to predict algal bloom. Water Sci. Technol. 36, 89–97.

Yamahara, K.M., Layton, B.A., Santoro, A.E., Boehm, A.B., 2007. Beach sands along the California coast are diffuse sources of fecal bacteria to coastal waters. Environ. Sci. Technol.