Self-Organizing Linear Output (SOLO) Approach for Managing Total Coliform Indicator Bacteria on California Beaches

J.-B. Kim†, H.-K. Bae†*, K.-L. Hsu‡, and S. Sorooshian‡

†Department of Global Environment, Keimyung University, Dalsegu
Shindangdong 1000, Daegu, 704-701, S. Korea
*e-mail: hunkyunbae@kmu.ac.kr

‡Department of Civil and Environmental Engineering,
University of California, Irvine, California, 92697, U.S.A

ABTRACT

Kim, J.-B., Bae, H.-K., Hsu, K.-L. and Sorooshian, S., 2011. Self-Organizing Linear Output (SOLO) Approach for Managing Total Coliform Indicator Bacteria on California Beaches, Journal of Coastal Research, SI 64 (Proceedings of the 11th International Coastal Symposium), 1063 -1067, Szczecin, Poland, ISSN 0749-0208

In this study, one of the artificial neural networks, Self-Organizing Linear Output (SOLO), was used to predict levels of indicator bacteria at Newport Bay in Newport, Beach, California, USA. The approach over-estimated several observations which showed extraordinary low concentrations compared to others. Average of observations was 6351 CFU/100mL and that of error value for model validations was only 176 CFU/100mL, about 3% of observation average without few points which showed extraordinary low concentrations. The results of this study showed that the approach was very effective for predicting concentrations of indicator bacteria. The approach was carried out for monthly average prediction because of limited dataset. The study could be extended for finer time scale prediction, such as weekly or daily prediction, when more measurements are available.

Keywords: A*rtificial Neural Network, Beach Pollution, Rainfall, Suspended Solids*

INTRODUCTION

The quality of water in coastal areas is a vital component to human activities as well as the natural ecosystem. Coastal areas are essential habitats for many species including threatened and endangered species. Population growth, however, along with increased urbanization in desirable coastal areas have resulted in increasing the amount of contaminants deposited in the ocean through the drainage system. This has negative impacts on coastal ecosystems, human beings, tourism and other economic activities. In this respect, water quality of the coastal region in California is extremely important since 80 percent of the state's population resides along the state's coastline and millions of tourists and local residents visit the coastal area in California each year, which demonstrates the importance of its coastal area to the economy and the culture of the state. Therefore, maintaining and monitoring water quality of these areas are a major challenge to the state of California because wrong decisions on water quality will greatly affect the health of coastal communities and their economy, respectively (SWRCB, 2001). To address this problem the state set up the water quality standard based on indicator bacteria concentration since bacterial concentration in beaches or surf zone is obviously the main concern to the State of California and its coastal water quality agencies or managers responsible for protecting beach-goers from exposure to waterborne disease.

The primary objective of this study is developing a modeling method to predict the level of water quality, especially bacterial concentration. There have been several efforts to develop the modeling method to forecast bacterial concentration (Auer and Niehaus, 1993; Sperling 1999; Steets and Holden, 2003; Reeves et al. 2004). Among those studies, two studies focused on Southern California. One, by Steets and Holden (2003), developed a mathematical model to predict fecal coliform concentration in the Arroyo Burro Lagoon in Santa Barbara, California and its adjacent coastline, Hendry's Beach. The other, by Reeves et al. (2004), developed the simple theoretical model to estimate the loading of fecal indicator bacteria in storm runoff originated from the sediment erosion in Talbert Watershed. Both studies adopted mathematical models using several parameters. In this study, however, we used an approach with an artificial neural network (ANN) to predict bacterial concentration. Many studies have already adopted ANNs to manage the water quality (Brion and Lingireddy 1999; Aguilera et al. 2001; Lou and Nakai, 2001; Cheroutre-Vialette and Lebert 2002; Ha and Stenstrom, 2003). Most of those studies used typical ANNs which have one or more hidden layers between input layer and output layer. This Study, however, adopted a unique ANNs, called Self-Organizing Linear Output (SOLO) which has classification and mapping layers instead of the hidden layers of typical ANNs (Hsu et al., 2002).

METHODS

Site Description

Both Newport Beach and Newport Bay areas are very popular places in Southern California. Newport Bay is the second largest estuarine embayment in Southern California. Pollution of the bay is largely dependent on the contaminants from the San Diego Creek Watershed, upper basin watershed, which covers 112.2 square miles in the center of Orange County, California (Kamer, 2002). As reported by Horne (2003), excess nitrate in San Diego Creek was likely the cause of eutrophication for the Newport Bay Estuary in 1993. The reason why the Newport Bay water quality situation in the 80's was worse than that of the 70's was because of the development of the city of Irvine, being the main part of

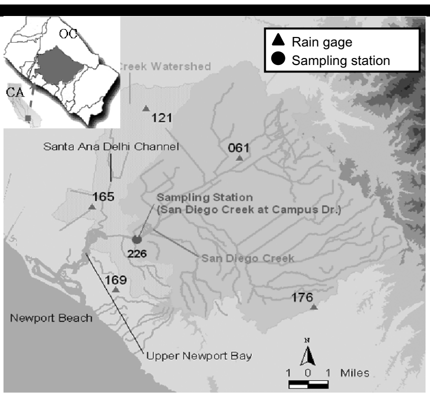

Figure 1 San Diego Creek Watershed and Study Area

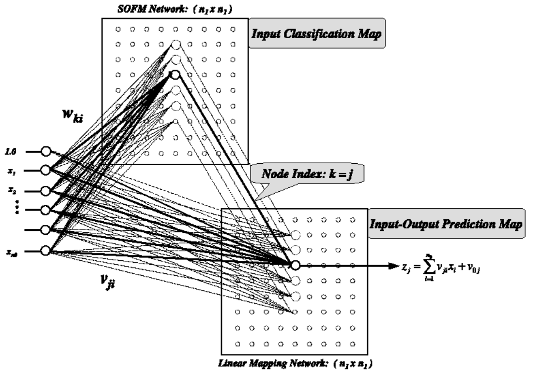
SOFM Network: (n1 x n1)
Input Classification Map
wki
1.0
x1
x2
xn0
Node Index: k = j
Input-Output Prediction Map
vji
$z_j = \sum_{i=1}^{n_0} v_{ji} x_i + v_{0j}$
Linear Mapping Network: (n1 x n1)

Figure 2 The architecture of a SOLO network

San Diego Creek Watershed. San Diego Creek, the main tributary of San Diego Creek Watershed, is the primary freshwater input to Upper Newport Bay and also the repository for agricultural and urban drainage throughout the watershed. For example, more than 95% of freshwater which flows into the Newport Bay comes from San Diego Creek. San Diego Creek is also the main contributor for the most pollutants from the San Diego Creek Watershed to Newport Bay; 95% of dissolved metals, approximately 94% of sediment (Strauss, 2002). Therefore, the present study focused on San Diego Creek and its upper streams flow. Figure 1 shows the study area. Triangles are rainfall gages; dot is the water quality monitoring point; and dark area is the study area. The effect of Santa Ana Delhi channel to the water qualities on Newport Bay is minor although it covers a Northwest part of San Diego Creek Watershed, so the area covered by Santa Ana Delhi was not considered for this study.

Data

Orange County Watershed and Coastal Resources Division provided dataset such as streamflow, precipitation, and water quality data. Water quality samples were taken at least once a week from the Station #226 (San Diego Creek at Campus Dr.) where streamflow is also recorded regularly. Precipitation data were recorded in real-time from five different gages as shown in Figure 1. The average precipitation was calculated using the thiessen polygon method. The Ocean Water Protection Program operated by Orange County Health Care Agency provided data for concentrations of total coliform (TC), one of the major indicator bacterial. Indicator bacteria samples were taken from the sampling station CNBCD, the same location as station #226.

Self-Organizing Linear Output

ANNs are the computational tool which mimics the biological processes of the human brain. Since the 1950's, many researchers have devoted themselves to study ANNs and multilayer feed-forward networks have been found to have the best performance with regard to input-output function among that research. SOLO is the one of the multivariate ANNs procedure with a classification and a mapping layer (Hsu et al., 2002). SOLO was designed for rapid, precise, and inexpensive estimation of network structure/parameters and system outputs. Moreover, SOLO provides features that facilitate insight into the underlying processes, thereby extending its usefulness beyond forecast applications. Figure 2 shows the structure of SOLO network. Input layer consists of n0 neural units connected to all units of the classification and mapping layers which categorizes the input variables into a certain number of characteristic groups, so each group could represent input data patterns and route these characterized input variables for output prediction; both layers are n1 x n1 matrixes. The classification layer uses a self-organizing feature map (SOFM) to classify the input information and mapping layer uses multivariate linear regression to map the inputs into the outputs (Hsu et al. 2002).

Experimental Setup

Providing effective input variables might be the most important task for modeling approaches. The study, therefore, focused on determining effective input variables, which could lead to improve model performance, at the beginning stage of ANN development. Several factors, turbidity, suspended solid (SS), precipitation, nutrient (N, P), and pH, which might affect bacterial concentrations were evaluated for input variable selection. Precipitation and SS showed the most highly relevant to the simulation results, so that those two parameters were used in the further experiment.

Model Scenario with Input – Output Pairs

Every combination of single or multiple input variables with different numbers of previous events and different node size from 2x2 to 7x7 was tested and each scenario contains previous indicator bacterial concentrations. The first 36 months data were used for training the model and the last 18 months data were used for verifying the model. Six months data were used for both training and verifying the model because of lack of dataset. Following input-output equations show each scenario; examples for single input variable (rainfall or SS) and for combination of input variables (rainfall and SS).

$$\begin{bmatrix} C^{sim}(2) \\ \mathrm{M} \\ C^{sim}(t) \end{bmatrix} = f \begin{bmatrix} I_1^{obs}(1) & C^{obs}(1) \\ \mathrm{M} & \mathrm{M} \\ I_1^{obs}(t-1) & C^{obs}(t-1) \end{bmatrix}$$

$$\begin{bmatrix} C^{sim}(3) \\ \mathrm{M} \\ C^{sim}(t) \end{bmatrix} = f \begin{bmatrix} I_1^{obs}(2) & I_1^{obs}(1) & C^{obs}(2) & C^{obs}(1) \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} \\ I_1^{obs}(t-1) & I_1^{obs}(t-2) & C^{obs}(t-1) & C^{obs}(t-2) \end{bmatrix}$$

$$\begin{bmatrix} C^{sim}(4) \\ \mathrm{M} \\ C^{sim}(t) \end{bmatrix} = f \begin{bmatrix} I_1^{obs}(3) & I_1^{obs}(2) & I_1^{obs}(1) & C^{obs}(3) & C^{obs}(2) & C^{obs}(1) \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} \\ I^{obs}(t-1) & I^{obs}(t-2) & I^{obs}(t-3) & C^{obs}(t-1) & C^{obs}(t-2) & C^{obs}(t-3) \end{bmatrix}$$

$$\begin{bmatrix} C^{sim}(4) \\ \mathrm{M} \\ C^{sim}(t) \end{bmatrix} = f \begin{bmatrix} I_1^{obs}(3) & I_1^{obs}(2) & I_1^{obs}(1) & I_2^{obs}(3) & I_2^{obs}(3) & I_2^{obs}(1) & C^{obs}(3) & C^{obs}(2) & C^{obs}(1) \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} \\ I_1^{obs}(t-1) & I_1^{obs}(t-2) & I_1^{obs}(t-3) & I_2^{obs}(t-3) & I_2^{obs}(t-3) & I_2^{obs}(t-3) & C^{obs}(t-1) & C^{obs}(t-2) & C^{obs}(t-3) \end{bmatrix}$$

Where C : concentration for indicator bacteria, TC,
I_i : input variable, rainfall or SS
t : time (month)

RESULTS and DISCUSSION

Input Variable: Precipitation

Rainfall events are the one of the most important factors for bacterial concentration because rainfall runoff washes all pollutants out to the water bodies from non-point sources. On the other hand, rainfall events could also dilute the bacterial concentrations because of its large water volume. In this study area, rainfall events tend to increase indicator bacterial concentrations. For example, there is a report which showed the effect of rainfall on the bacterial concentrations in Southern California. Noble et al. (2003) reported bacterial concentration changes of three different seasons along the Southern California Coastline. The concentration of indicator bacteria in the coastal zones is 5 to 40 times higher during the storm event than those during any other seasons in the same year. Another report showed the strong relationship between rainfall events and bacterial concentrations in a different area, Stevenson Creek Watershed in Florida. Whitlock et al. (2002) reported that the changes of bacterial concentrations showed the same patterns with those of rainfall events in Stevenson Creek Watershed. Figure 3 (a) and (c) show average rainfall for the study area and average of TC concentrations measured from the San Diego Creek at Campus Dr. station. As shown in the figure, there is a strong positive relationship between rainfall and TC concentration. Bacterial concentrations might be affected by other factors during dry season, so the study needed to have other forcing data to cover dry season.

Input Variable: Suspended Solids (SS)

SS usually indicates particles between 0.1mm and 2 mm in the liquid sample. SS is also a possible candidate as an input variable for modeling approach to predict bacterial concentrations because bacteria could attach to sediment particles. For instance, Auer and Niehaus (1993) reported that 90 ~96% of fecal coliform were associated with particles between 0.45 and 10 and Steets et al. (2003) reported that 90% of fecal coliform in streams and bays are associated with sediments. Kimberly et al. (2005) also reported that fecal coliform in sediments could live longer than those in water. Figure 3 (b) and (c) show averages of SS and TC concentrations collected at the sampling station at Campus Dr. Changes of particles also showed strong relationship with those of bacteria. Changes of SS also explained wash-out effects with the first rainfall. SS concentrations didn't show proportional relationship with amount of rainfall because pollutants were usually washed out with the first rainfall event, so concentrations of pollutants tend to be decreased during following rainfall events. Changes of SS for the study area showed this fact well.

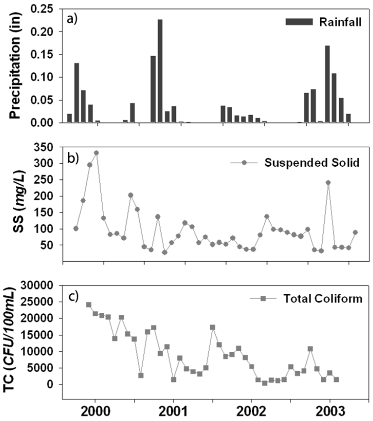

Figure 3 Monthly Average Data

(a) Average Rainfall in study area

(b) SS in San Diego Creek at Campus Dr. Station

(c) TC concentrations in San Diego Creek at Campus Dr. Station

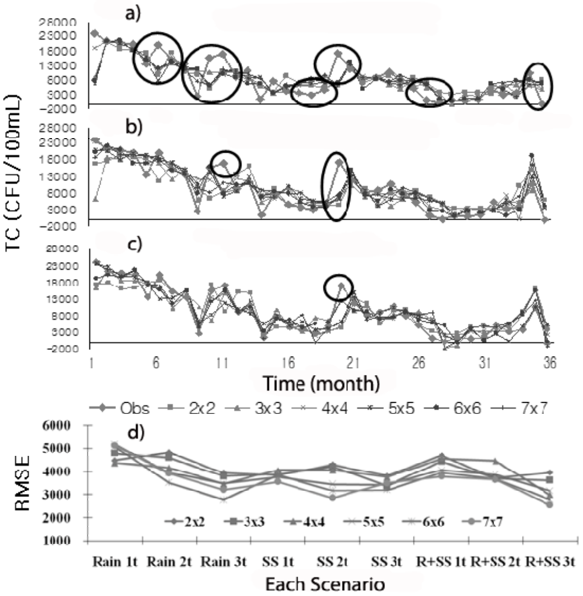

Figure 4 Calibration Result for TC Concentration prediction with Rainfall as Input Variable and RMSE for all calibration cases

a) One step previous rainfall data as input variable

b) Two steps previous rainfall data as input variable

c) Three steps previous rainfall data as input variable

d) RMSE for all cases

Model Calibration

Figure 4 a), b), and c) show calibration results for predicting TC concentrations from one of model scenarios, single input variable (rainfall). X axis represents time, y axis TC concentration, line with diamond symbol observations, and other lines and shapes simulation results with different node sizes. Figure d) shows the root mean squared error (RMSE) for all simulations; x axis each scenario, y axis RMSE values. Calibration results showed that the model performance depended on how many previous data was considered as input variables. The more previous data, the better simulation results. The scenario with one step previous rainfall data as input variable showed several points which over- or under-estimated to the observations (circles or ovals in Figure 4 a)) while other twoinput scenarios had only a few under-estimations. The node size did not affect model performance when rainfall was the only input. Figure 4 d) shows all calibration results in terms of error value, RMSE. As shown in the figure, the most complicated input scenario, combined parameters - rainfall and SS, three previous events of input variables, and 7x7 nodes, had the best result and simpler one, one time previous rainfall data with 7x7 nodes, the worst. There is no significant difference among each node size since each scenario had own node size for the best result. However, bigger node sizes, 6x6 and 7x7, had slightly better results than other node sizes in most cases. Overall, the more input scenario with bigger node size had better performance for calibration processes.

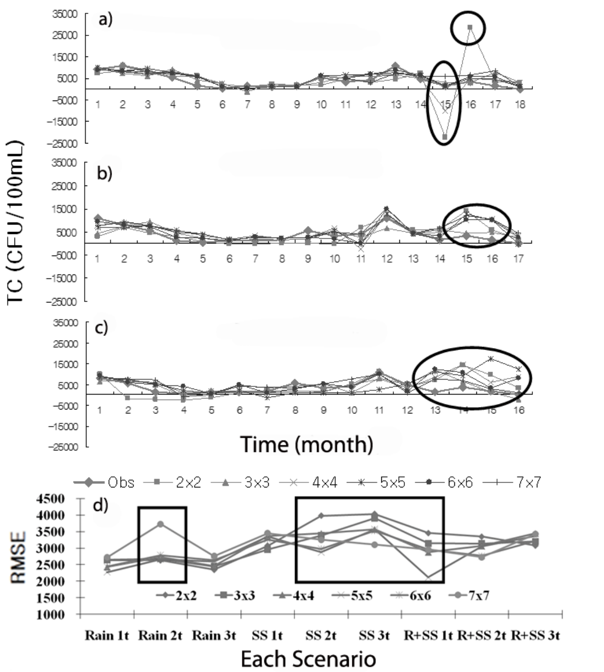

Figure 5 Validation Result for TC Concentration prediction with Rainfall as Input Variable and RMSE for all validation cases

a) One step previous rainfall data as input variable

b) Two steps previous rainfall data as input variable

c) Three steps previous rainfall data as input variable

d) RMSE for all cases

Model Validation

Figure 5 a), b), and c) show validation results for TC concentration predictions from the scenario which rainfall was the only input. The best performance was shown with one step previous rainfall data, unlikely the most complicated case showed the best results for calibration process. The scenario with one step previous rainfall data and 2x2 node size as well as 4x4 node size showed two under-estimations and one over-estimation (circle and oval in Figure 5 a)), but others followed observations well. Validations with two steps previous rainfall data and three steps previous rainfall data captured observations well at the early parts of simulations, but had over-estimations for the later parts (ovals in Figure 5 b) and c)). Figure 5 d) shows the changes of RMSE for validations with all scenarios. In terms of objective function, validation results showed different performance dependent on input scenarios (rectangular in Figure 6 d)). Estimations with rainfall as single input variable had better results than those with other cases in general, although the best result was shown in the scenario with multiple input variables, rainfall and SS, one step previous input data and node size 5x5. The average of error value for validation processes was 3041 CFU/100mL and the average of observation was 4990 CFU/100mL, so model performance for validation was poor if considering only error value. Several extraordinary low points of observations, which concentrations showed less than 1500 CFU/100mL, brought this poor estimation. Without those extraordinary low points, the average of observations was 6351 CFU/100mL and that of error value was only 176 CFU/100mL which is about 3% of observations average.

Overall, Rainfall and SS were the most important factors to predict bacterial concentrations using the SOLO approach. For the calibration, the more input variables, the better predictions and the node size didn't affect model performance. For the validation, model performance had slightly different result dependent on node size as well as input variables. The best result was shown in the scenario with multiple input variable, rainfall and SS, one step previous input data and node size 5x5 in terms of objective function.

CONCLUSION

The study focused on a SOLO approach to predict TC, one of the major indicator bacteria. Observation data had few points which were relatively lower than others and SOLO estimations couldn't capture those lower peaks. Although SOLO missed those lower concentrations of observations, SOLO seemed to have a predictable aptitude for high levels of concentrations. This fact is very important since high levels of bacterial concentrations are more critical regarding water quality issues. In this respect, the SOLO approach could be considered as a promising method to predict bacterial concentrations in beach areas. Beach pollution has already become a serious problem in the State of California and has created events that have affected the state economy as well as the culture of coastal areas. The state regulates the quality of recreational use water based on indicator bacteria, but the system could not give the information for the conditions of water quality in a timely manner. A modeling approach may be the one of options to use along with the current monitoring system since now- or fore-casting events could be possible. The study showed the possibility to use SOLO as a prediction tool for the bacterial concentrations. However, the study was limited to monthly event predictions because of the lack of dataset, so further studies for finer time scale prediction, such as weekly or daily changes of indicator bacterial concentration, as well as finding more variables which could produce better model performance are required.

REFERENCE

Aguilera, P.A.; Frenich, A.G.; Torres, J.A.; Castro, H.; Martinez Vidal, J.L. and Canton, M., 2001. Application of the Kohonen Neural Network in coastal water management: Methodological development for the assessment and prediction of water quality, *Wat. Res*., 35(17), 4053-4062

Auer, M.T. and Niehaus, S.L., 1993. Modeling fecal coliform bacteria - I. Field and laboratory determination of loss kinetics, *Wat. Res*., 27(4), 693-701

Brion, G.M. and Lingireddy, S., 1999. A neural network approach to identifying non-point sources of microbial contamination, *Wat. Res*., 33(14), 3099-3106

California Regional Water Quality Control Board, 1999. Amendment to the basin Plan, Santa Ana Region, Chapter 5 - Implementation Plan, Discussion of Newport Bay Watershed, California, USA

Cheroutre-Vialette, M. and Lebert, A., 2002. Application of recurrent neural network to predict bacterial growth in dynamic conditions, International Journal of Food Microbiology, 73(2-3), 107-118

French, C.B., 2003. Modeling Nitrogen Transport in the Newport Bay/San Diego Creek Watershed, California: University of California, Riverside, Master's degree Thesis

Ha, H. and Stenstrom, M.K., 2003. Identification of land use with water quality data in stormwater using a neural network, *Wat. Res*., 37(17), 4222 - 4230

Horne, A.J., 2003. Eutrophication in the Newport Bay-Estuary in 2002: Trends in the Abundance of Nuisance Macroalgae (Seaweed) in 1996-2002, Report to Orange County Public Facilities & Resources Department, California, USA

Hsu, K.-L.; Gupta, H.V.; Gao, X.; Sorooshian, S. and Imam, B.,

2002. Self-organizing linear output map (SOLO): An artificial neural network suitable for hydrologic modeling and analysis, *Wat. Res. Res*., 38(12), 1302-1319

Kamer, K.; Schiff, K.; Kennison, R.L. and Fong, P., 2002. Macroalgal Nutrient Dynamics in Upper Newport Bay, Technical Report, Southern California Coastal Water Research Project, California, USA

Kimberly L.A.; John E.W. and Valerie J. H., 2005. Persistence and Differential Survival of Fecal Indicator Bacteria in Subtropical Waters and Sediments, *Appl. Environ. Micro*., 71(6), 3041-3048

Lou, W. and Nakai, S., 2001. Application of artificial neural networks for predicting the thermal inactivation of bacteria: a combined effect of temperature, pH and water activity, *Food Research International*, 34(7), 573-579

Noble, R.T.; Moore, D.F.; Leecaster, M.K.; McGee, C.D. and Weisberg, S.B., 2003. Comparison of total coliform, fecal coliform, and enterococcus bacterial indicator response for ocean recreational water quality testing, *Wat. Res*., 37(7), 1637-1643

Reeves, R.L.; Grant, S.B.; Mrse, R.D.; Copil Oancea, C.M.; Sanders, B.F. and Boehm, A.B., 2004. Scaling and Management of Fecal Indicator Bacteria in Runoff from a Coastal Urban Watershed in Southern California, *Environ. Sci. Technol*., 38(9), 2637- 2648

Sperling, M.V., 1999. Performance evaluation and mathematical modeling of colifrom die-off in tropical and subtropical waste stabilization ponds, *Wat. Res*., 33(6), 1435-1448

Steets, B.M. and Holden, P.A., 2003. A mechanistic model of runoff-associated fecal coliform fate and transport through a coastal logoon, *Wat. Res*., 37(3), 589- 608

Strauss, A., 2002. Total Maximum Daily Loads for toxic pollutants San Diego Creek and Newport Bay, California, U.S. Environmental Protection Agency Region 9, California, USA

SWRCB (State Water Resources Control Board, California Environmental Protection Agency), 2001. Source investigations of storm drain discharges causing exceedances of bacteriological standards, California, USA

Whitlock, J.E.; Jones, D.T. and Harwood, V. J., 2002. Identification of sources of fecal coliforms in an urban watershed using antibiotic resistance analysis, Wat. Res., 36(17), 4273-4282