Scientia Iranica, Vol. 15, No. 3, pp 308–314

SCIENTIA IRANICA

Use of Artificial Neural Networks in Predicting Highway Runoff Constituent Event Mean Concentration

A. Massoudieh[1] and M. Kayhanian*

In this paper, the large amount of highway runoff characterization data that were collected in California, during a 3-year monitoring season (2000-2003), were assessed in order to develop an Artificial Neural Network (ANN) model for predicting the Event Mean Concentration (EMC) of the constituent. The initial data analysis performed by a Multiple Linear Regression (MLR) model revealed that the Total Event Rainfall (TER), the Cumulative Seasonal Rainfall (CSR), the Antecedent Dry Period (ADP), the contributing Drainage Area (DA) and the Annual Average Daily Traffic (AADT) were among the variables having a significant impact on the highway runoff constituent EMC. These parameters were used as the basis for developing an Artificial Neural Network (ANN) model. The ANN model was also used to evaluate the impact of various site and storm event variables on highway runoff constituents' EMCs. The ANN model has proven to be superior to the previously developed MLR model, with an improved R^2 for most constituents. Through the ANN model, one was able to see some non-linear effects of multi variables on pollutant concentration that, otherwise, would not have been possible with a typical MLR model. For example, the results showed that copper EMC is more sensitive at higher Annual Average Daily Traffic (AADT), with respect to ADP, compared with lower range AADT.

INTRODUCTION

Storm water has recently been considered a major pollution source of many urban waters [1-5] and highway runoff has received much attention due to the appearance of heavy metals, hydrocarbons and fuel additives [6-8]. Some researchers have specifically concentrated on evaluating the characteristics of runoff, based on particle size distribution and pollutant adsorption [9-12]. Most of these studies were, however, performed locally and there is a lack of information on larger watersheds or on a statewide basis, with an emphasis on parameters that influence the runoff quality. Only few studies have attempted to develop a relationship between pollutants and some observed variables using Multiple Linear Regressions (MLR). For example, Irish et al. [13] and Wu et al. [14] used regression modeling to analyze highway storm water loads. A few studies tried to investigate the possible relationship between pollutant load and variables (e.g., annual average daily traffic), measured during the storm events [15-16]. Literature reviewed by Wistrom and Matsomoto [17], however, concluded that AADT is not generally expected to be useful as a control variable for the design, operation and maintenance of specific runoff control structures, as traffic intensity on a particular stretch of highway is expected to be fairly constant from day to day. While, in some cases, a good correlation coefficient was obtained, these researchers caution that predictive equations may not be useful outside the study regions.

In recent years, an Artificial Neural Network (ANN) has also been used for exploring the relationships between the input and output of environmental conditions, including water resources and environmental engineering. ANN is a computational approach using the theories of massive interconnected and parallel processing of biological systems. The massively

1. *Department of Civil and Environmental Engineering, University of California, Davis, CA 95616, USA.*

*. *Corresponding Author, Center for Environmental and Water Resources Engineering, Department of Civil and Environmental Engineering, One Shields Avenue, Engineering III, University of California, Davis, CA 95616, USA. E-mail: mdkayhanian@ucdavis.edu*

parallel distributed structure of ANN and its ability to explore non-linear relationships between inputs and outputs make it a suitable method for complicated non-linear modeling problems [18]. Some examples of recent ANN applications include rainfall runoff modeling [19], wastewater treatment [20], ecological modeling in aquatic environments [21], dispersion of atmospheric pollutants [22], land use and water quality [23], aquifer parameter estimation [24] and stormwater utility [25]. In this study, a large highway runoff quality data set has been used that was collected throughout the state of California in order to (1) use ANN to evaluate the interdependent relationship between the environmental variables and constituent EMCs and (2) explore the capability of ANN in estimating the constituents' EMC.

METHODS

Data Collection

Storm water runoff characterization data used in this study were obtained from over fifty highway sites in California covering a wide range of annual average daily traffic levels and environmental conditions [8,26]. As part of this comprehensive monitoring program, automatic equipment was used to collect flow-weighted composite samples, measure runoff flow rate and rainfall amounts. On average, up to eight storm events were monitored annually at each highway site during the wet seasons (October through April) for a 3-year (2000-2003) period. Depending on the storm intensity and duration, up to 50 aliquot samples were obtained to capture a representative composite sample during each event.

The flow-weighted composite samples obtained from the entire storm event were sent to a laboratory for analysis. The results of these analytical tests are assumed to represent Event Mean Concentrations (EMC) for runoff from a given rainfall event. Constituents and parameters routinely analyzed under this program include conventionals (pH, temperature, hardness, conductivity), aggregate (O&G, TSS, TDS, TOC and DOC), metals (total and dissolved As, Cd, Cr, Cu, Pb, Ni and Zn) and nutrients (nitrate, TKN, total and dissolved P).

Extensive field and laboratory Quality Assurance/Quality Control (QA/QC) procedures were followed. Analytical results were qualified as necessary, based on the results of the QA/QC evaluations using newly developed data validation software as described in Kayhanian [27]. Chemical constituents containing data below a detection limit were analyzed using regression-on-order statistics described in Shumway et al. [28]. A routinely analyzed water quality data and statistical summary report is presented in Table 1. Additional detailed information on Caltrans storm water runoff characterization studies, including the monitoring site locations and automatic sampling equipment, can be obtained from Kayhanian et al. [8].

The validated data were then imported into an access database containing three main tables: Sample description, event description and site description. The database was used to extract all analytical data, precipitation information and site characteristics data for statistical analysis and ANN model development.

ARTIFICIAL NEURAL NETWORK (ANN) MODEL

ANN analysis was performed (1) To evaluate the interdependent characteristics of site and rainfall on constituents' EMC and (2) To explore its application as a predictive tool to estimate the highway runoff constituents' EMC. Wide forms of ANN have been applied in the field of science and engineering. The most common form of ANN is the feed forward ANN. A typical feed forward ANN includes an input layer, an output layer and one or more hidden layers. Each of these layers contains several nodes or neurons, which are connected to each other by some multiplication factors called synaptic weights. Each neuron contains a function called an "activation function" that can be a step, linear or non-linear function. The outputs of each layer are used as the input for the consequent layer. For each constituent, various configurations of ANN (e.g., different numbers of hidden layers and neurons in each layer) were evaluated (i.e., R^2 and Standard Error (SE) for estimation and validation subsets) in order to find the most efficient configuration of the network with the most acceptable generalization capability. Several degrees of complexity for selecting the number of nodes and layers were tested to determine the most appropriate network configuration for each constituent that could provide the best prediction, as well as generalized capability. The result of this analysis for total copper is shown in Figure 1. As can be noted, the R^2 for estimated values greatly improved, as the number of nodes increased. However, the model loses its generalization capability by choosing a high level of complexity (i.e., R^2 for validation data decreased). To achieve a reliable level of generalized R^2 , the values obtained for the estimation data should not be significantly larger than for the validation data. In case of total copper, a network with 3 nodes in one hidden layer was found to provide the best prediction efficiency as well as generalization capability. The optimum node for most other constituents was found to be the same.

After finding the optimum node in the hidden layer, the most useful and effective (optimum) input variables were determined. Optimum input variables

Table 1. Statistical summary of highway runoff quality in California during 2000-2003.

Constituent	Unit	Summary Statistics					
		N	% Detect[a]	Range	Mean	Median	SD
Conventional							
EC	μs/cm	634	100	5-743	96.1	72.7	73.4
Hardness	mg/L as $CaCO_3$	635	99	2-400	36.5	26.9	34.2
pH	pH unit	633	100	4.5-10.1	7.1	7	0.7
Temp.	°C	183	100	4.7-25.4	12.5	12	3.4
Aggregate							
DOC	mg/L	635	100	1.2-483	18.7	13.1	26.2
O&G	mg/L	39	70	1-20	6.6	6	4.2
TDS	mg/L	635	97	3.7-1800	87.3	60.3	103.7
TOC	mg/L	635	100	1.6-530	21.8	15.3	29.2
TSS	mg/L	634	100	1-2988	112.7	59.1	188.8
Metals (Dissolved)							
As	μg/L	635	40	0.5-20	1.0	0.7	1.4
Cd	μg/L	635	42	0.2-8.4	0.24	0.13	0.5
Cr	μg/L	635	80	1-23	3.3	2.2	3.3
Cu	μg/L	635	100	1.1-130	14.9	10.2	14.4
Ni	μg/L	635	79	1.1-40	4.9	3.4	5.0
Pb	μg/L	635	60	1-480	7.6	1.2	34.3
Zn	μg/L	635	99	3-1017	68.8	40.4	96.6
Metals (Total)							
As	μg/L	635	62	0.5-70	2.7	1.1	7.9
Cd	μg/L	635	76	0.2-30	0.7	0.44	1.6
Cr	μg/L	635	97	1-94	8.6	5.8	9.0
Cu	μg/L	635	100	1.2-270	33.5	21.1	31.6
Ni	μg/L	635	95	1.1-130	11.2	7.7	13.2
Pb	μg/L	635	94	1-2600	47.8	12.7	151.3
Zn	μg/L	635	100	5.5-1680	187.1	111.2	199.8
Nutrients							
NO_3-N	mg/L	634	90	0.01-48	1.07	0.6	2.4
Ortho-P	mg/L	630	64	0.01-2.4	0.11	0.06	0.2
Total P	mg/L	631	89	0.03-4.69	0.29	0.18	0.4
TKN	mg/L	626	94	0.1-17.7	2.06	1.4	1.9

N = Sample size, SD = Standard deviation,

[a] = Values below detection limit is denoted as "non-detect". For constituents containing non-detects, the statistical method outlined by Shumway et al. [28] was used to determine summary statistics.

were determined by including all measured site and rainfall characteristics, which included Drainage Area (DA), impervious fraction, landuse, Annual Average Daily Traffic (AADT), Total Event Rainfall (TER), Seasonal Cumulative Rainfall (SER), rain intensity and Antecedent Dry Period (ADP). After reviewing the results of this analysis, it was concluded that five variables, including TER, SER, ADP, DA and AADT, are predominant parameters influencing most constituents' EMC. For this reason, the mentioned five parameters were used as input variables for the ANN model. A schematic of the ANN with two hidden layers and five input variables used in this study is shown in Figure 2. For the ANN shown in Figure 2, the expression describing the relationship between the input and output for ANN with 2 hidden layers can be

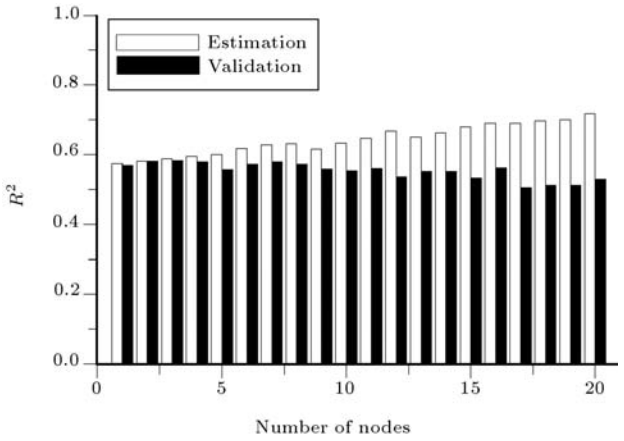

Figure 1. Optimum ANN model for total Cu based on R^2 and number of nodes in the hidden layer.

written, using the following matrix notation:

$$\text{Output} = W_3.\psi[W_2.\psi(W_1.\text{Input} + B_1) + B_2] + B_3, \quad (1)$$

where:

$W_k = [w_{i,j,k}]_{n_k \times n_{k-1}} =$ matrix of synaptic weight connecting layer $k-1$ to layer k,

$B_k = [b_{ik}]_{n_k} =$ bias vector for layer k,

$\psi =$ the activation function.

To use ANN as a predictive tool, it was first trained. The goal in training an ANN is to find the synoptic weights (W) and bias vectors (B) that can best represent or predict the relationship between the inputs and outputs by minimizing the difference between the predicted values by the network and the observed values. The normalized natural logarithms of the input variables and the target value were used for training. As an example, the representative value for AADT was computed using the following expressions:

$$\overline{\text{AADT}} = \frac{\ln(\text{AADT}) - \text{mean}[\ln(\text{AADT})]}{\sqrt{\text{Var}[\ln(\text{AADT})]}}. \quad (2)$$

An artificial neural network is said to generalize well when the input-output mapping computed by the network is correct (or nearly so) for test data that was never used in creating or training the network [18]. One way to predict the outputs more precisely is by increasing the number of nodes and hidden layers. However, the prediction efficiency for the data that has not been introduced to the model may become poor. In this study, the optimum number of nodes and layers that could provide the most efficient prediction, while maintaining the generalization capability of the model, was determined.

The ANN toolbox of Matlab 6.0 software package (Wolfram research) was used for the training and validation of the network. In order to test the generalization capability of the model, the monitoring data was divided into two subsets: (i) An estimation subset, which was randomly selected and accounts for two-thirds of the whole data set and which was used for

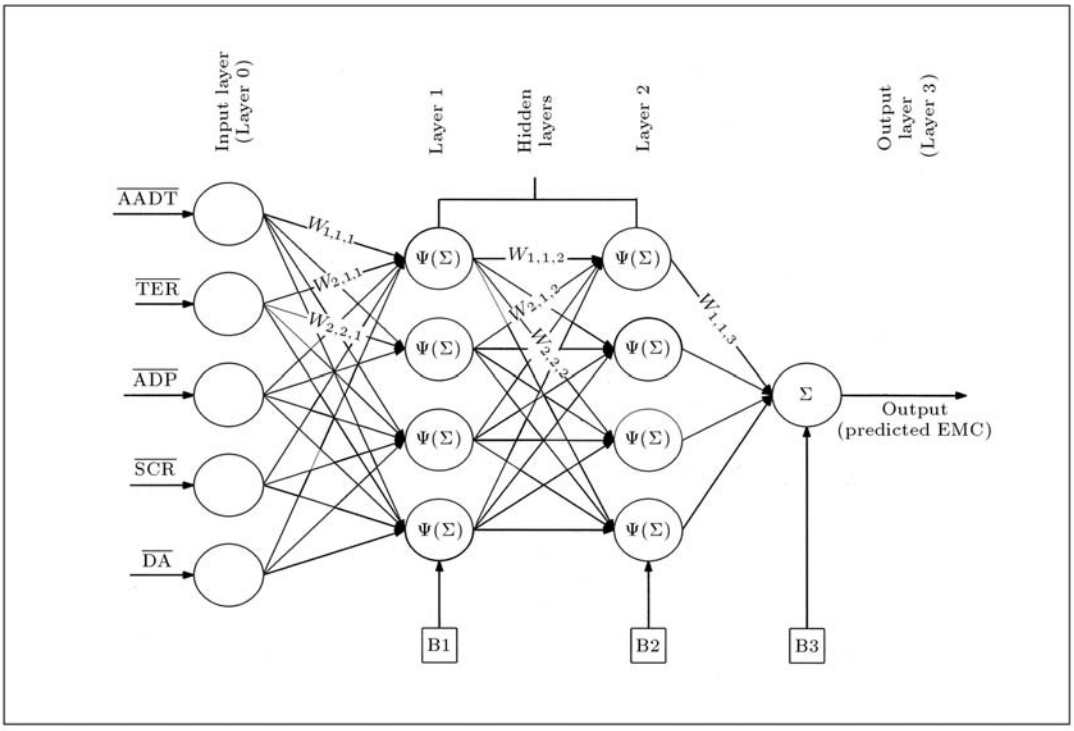

Figure 2. Schematic for an ANN with two hidden layers.

training the model and, (ii) A validation subset, which was the remaining data and which was used for testing the generalization capacity of the model.

RESULTS AND DISCUSSION

The influences of the predictor parameter on the pollutant event mean concentration and their interdependent variability were first examined by ANN analysis. The interdependent variability is an added advantage of using the ANN model, which, otherwise, would be more difficult or impossible to evaluate with an MLR analysis. Figure 3 present the relationship between different predictor variables and their interactions on EMC values for copper. For producing this figure, two of the predictor variables are chosen to vary within their ranges and the rest of the variables are kept constant and equal to their geometrical mean. The choice of the geometric mean as a representative for the mean of the values is assumed to be reasonable, since most of the predictor variables follow log-normal distribution. As can be seen from Figure 3, ANN is capable of showing the non-linear behavior that exists between the predictor values and EMCs. For example, it can be noticed that, for lower ranges of AADT (e.g. 1000-10,000), copper EMC does not vary significantly with an increase in ADP, while for higher ranges of AADT (60,000-200,000), EMC increases significantly with an increase of ADP. Similar behavior can be observed for TER with a strong negative relationship between EMC and TER at larger AADT values and a weaker correlation at smaller AADP values.

The agreement between the predicted ANN model, the observed concentrations of DOC and the total Cu for both estimation and validation subsets are shown in Figure 4. Due to the complexity of the relationships generated by the ANN method, it is not possible to present them in an explicit form (e.g. in form of an equation). However, a user friendly tool can be developed for an estimation of EMC values by the ANN method using Microsoft Visual Basic.

Constituent EMC predicted by the ANN model was compared with the prediction values, using the previous MLR model developed by Kayhanian et al. [29]. Comparison of MLR with the ANN model was mostly evaluated through R^2 values. For consistency in this comparative analysis, the same five variables, including AADT, TER, ADP, SCR and DA, were considered as the input variables for both ANN and MLR models. The R^2 values for both models, for selective constituents, are shown in Table 2. As can be noted, for most constituents, the R^2 values obtained from the ANN model have been shown to be superior to the R^2 obtained from the MLR model. For instance, the improvement made in ANN model R^2 for DOC and total Cu, relative to the R^2 found in the MLR

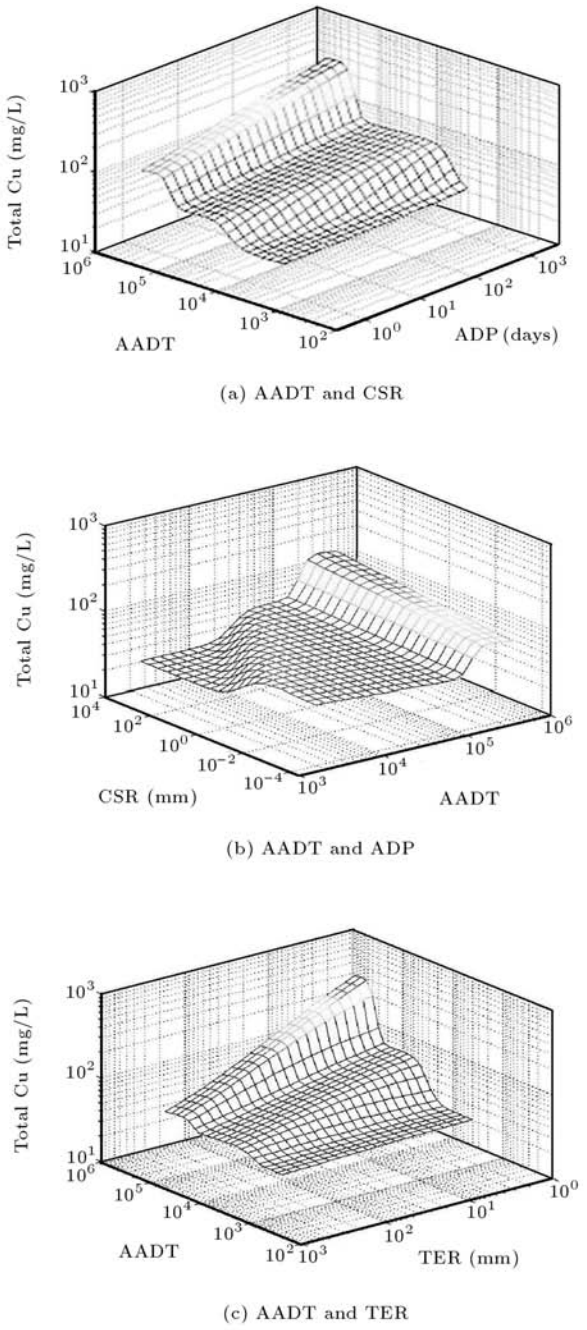

Figure 3. Variation of total Cu event mean concentration versus predictive variables.

model, is about 28 and 34 percent, respectively. For some constituents (e.g., dissolved Pb and Ortho-P), the improvement in R^2 was over 100%.

The ANN analysis has been able to successfully identify the environmental and site-specific factors that significantly affect the runoff quality. Knowledge of these factors and their effects on the runoff quality is useful in evaluating storm water management issues, such as in planning future monitoring efforts and in designing studies of pollutant removal effectiveness. The ANN model developed herein has been shown to

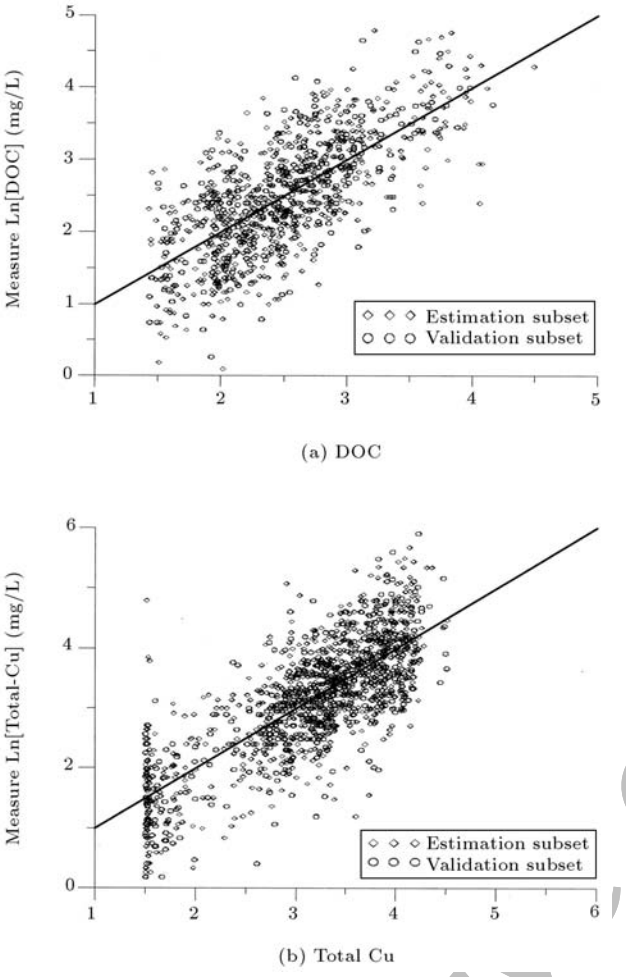

Figure 4. ANN model performance based on measured and predicted values.

predict the EMCs of certain constituents better than others for specific sites and storm events. The ANN model can also be used to provide improved estimates of long-term average concentrations or loads from highway facilities as a whole. Developing ANN models for runoff quality has a number of other practical applications. For instance, the modeling of runoff quality within a watershed allows better comparisons with relevant water quality discharge limits rather than the simple statistical estimates of percent exceedance. Additionally, neural networks can be used to determine the weighted averaged EMC from a sub-watershed to a better estimate of the constituent's load on a watershed basis. Based on the findings of this study, the ANN model may be used as a predictive tool for computing the constituent EMC, as well as being used as a "big picture" management decisions tool.

CONCLUSIONS

Major conclusions drawn from this study include:

Table 2. Comparison of R^2 values for MLR and ANN models.

Constituent	MLR R^2	ANN R^2	Percent Improvement
DOC	0.41	0.52	27%
TSS	0.25	0.44	76%
Cu-d	0.51	0.60	18%
Cu-t	0.52	0.70	35%
Pb-d	0.08	0.39	388%
Pb-t	0.36	0.69	92%
Zn-d	0.32	0.41	28%
Zn-t	0.51	0.63	24%
NO_3-N	0.37	0.47	27%
Ortho-P	0.15	0.38	153%
P-t	0.10	0.20	100%
TKN	0.39	0.41	5%

t = total and d = dissolved; example: Cu-d = dissolved Cu

1. Through the ANN analysis, the sensitivity of constituent's EMC, with respect to a predictor variable, has been examined. For example, for copper, the sensitivity with respect to ADP is high for AADT = 60,000 to 200,000, whereas, it may not be as sensitive or less sensitive in other ranges of AADT = 1,000-10,000;
2. The performance of ANN model prediction against measured values was evaluated with R^2. The ANN analysis performed in this study resulted in superior R^2 values compared to the previously developed MLR model. In general, R^2 improvements in the range of 7 to over 100 percent were observed;
3. A relatively good agreement between the ANN model prediction and the measured concentration of most constituents was observed. This good correlation could not be presented through an explicit equation form. However, a user friendly tool can be developed for estimation of EMC values by the ANN method using Microsoft Visual Basic;
4. Aside from constituent EMC estimation, AADT and other variables evaluated as part of the ANN analysis can also be used as follows: As a predictive planning tool for decision-making and load modeling and for prioritizing efforts for managing the runoff quality in highly urbanized area.

ACKNOWLEDGMENTS

The funding for the storm water runoff characterization studies used in this study was provided by the Division of Environmental Analysis (DEA), California Department of Transportation (Caltrans).

REFERENCES

1. Characklis, G.W. and Wiesner, M.R. "Particles, metals and water quality in runoff from large urban watershed", *J. Envir. Engrg., ASCE*, **123**(8), pp 753-759 (1997).
2. Barrett, M.E., Irish, L.B., Malina, J.F. and Charbeneau, R.J. "Characterization of highway runoff in Austin, Texas area", *J. Envir. Engrg., ASCE*, **124**(2), pp 131-137 (1995).
3. Davis, A.P., Shokouhian, M. and Ni, S.B. "Loading estimates of lead, copper, cadmium and zinc in urban runoff from specific sources", *Chemosphere*, **44**(5), pp 997-1009 (2001).
4. German, J. and Svensson, G. "Metal content and particle size distribution of street sediments and street sweeping waste", *Wat. Sci. Tech.*, **46**(6-7), pp 191-198 (2002).
5. Vaze, J. and Chiew, F.H.S. "Nutrient loads associated with different sediment sizes in urban stormwater and surface pollutants", *J. Envir. Engrg.*, **130**(4), pp 391-396 (2004).
6. Furumai, H., Balmer, H. and Boller, M. "Dynamic behavior of suspended pollutants and particle size distribution in highway runoff", *Wat. Sci. Tech.*, **46**(11-12), pp 413-418 (2002).
7. Khan, S., Lau, S.L., Kayhanian, M. and Stenstrom, M.K. "Oil and grease measurement in highway runoff-sampling time and event mean concentrations", *J. Envir. Engrg.*, **132**(3), pp 415-422 (2006).
8. Kayhanian, M., Suverkropp, C., Ruby, A. and Tsay, K. "Characterization and prediction of highway stormwater pollutant event mean concentrations", *Environmental Management*, **85**(2), pp 279-295 (2007).
9. Sansalone, J.J. and Buchberger, S.G. "Characterization of solid and metal element distributions in urban highway stormwater", *Wat. Sci. Tech.*, **36**(8-9), pp 155-160 (1997).
10. Sansalone, J.J., Koran, J.M., Smithson, J.A. and Buchberger, S.G. "Physical characteristics of urban roadway solids transported during rain events", *J. Envir. Engrg.*, **124**(5), pp 427-440 (1998).
11. Sansalone, J.J. and Tribouillard, T. "Variation in characteristics of abraded roadway particles as a function of particle size", *Transportation Research Record*, **1690**, pp 153-163 (1999).
12. Li, Y., Lau, S.-L., Kayhanian, M. and Stenstrom, M.K. "Particle size distribution in highway runoff", *J. Envir. Engrg.*, **131**(9), pp 1267-1276 (2005).
13. Irish, L.B., Barrett, M.E., Malina, J.F. and Charbeneau, R.J. "Use of regression models for analyzing highway storm-water loads", *J. Envir. Engrg.*, **124**(10), pp 987-993 (1998).
14. Wu, J.S., Allan, C.J., Saunders, W.L. and Evett, J.B. "Characterization and pollutant loading estimation for highway runoff", *J. Envir. Engrg.*, **124**(8), 584-592 (1998).
15. Kerri, K.D., Racin, J.A. and Howell, R.B. "Forecasting pollutant loads from highway runoff", *Transportation Research Record*, **1017**, pp 39-46 (1985).
16. Chui, T.W., Mar, B.W., Horner, R.R. "Pollutant loading model for highway runoff", *J. Envir. Engrg.*, **108**(6), pp 1193-1210 (1982).
17. Wistrom, A.O. and Matsumoto, M.R. "Highway runoff: contaminant sources and deposition mechanisms", Department of Chemical and Environmental Engineering, University of California, Riverside, USA (1999).
18. Haykin, S., *Neural Networks, a Comprehensive Foundation*, Prentice Hall, NJ (1999).
19. Torfs, P. and Wojcik, R. "Local probabilistic neural networks in hydrology", *Phys. Chem. Earth*, **26**(1), pp 9-14 (2001).
20. Moreno, A.N., Redondo, C.A.F. "Intelligent wastewater treatment with neural networks", *Water Policy*, **3**, pp 267-271 (2001).
21. Wilson, H. and Recknagel, F. "Towards a generic artificial neural network model for dynamic prediction of algal abundance in freshwater lakes", *Ecological Modeling*, **146**, pp 69-84 (2001).
22. Podnar, D., Koračin, D. and Panorska, A. "Application of artificial neural networks to modeling the transport and dispersion of tracers in complex terrain", *Atmospheric Environment*, **36**(3), pp 561-570 (2002).
23. Ha, H. and Stenstrom, M.K. "Identification of land use with water quality data in stormwater using a neural network", *Water Research*, **37**(10), pp 4222-4232 (2003).
24. Balkhair, S.K. "Aquifer parameters determination for large diameter wells using neural network approach", *Journal of Hydrology*, **265**, pp 118-128 (2002).
25. Lee, H. and Stenstrom, M.K. "Stormwater monitoring utility", *Water Research*, **77**(1), pp 23-33 (2004).
26. Kayhanian, M., Singh, A., Suverkropp, C. and Borroum, S. "Impact of annual average daily traffic on highway runoff pollutant concentrations", *J. Envir. Engrg.*, **129**(9), pp 975-990 (2003).
27. Kayhanian, M. "Advanced stormwater runoff characterization", *Proceedings of the 10th International Conference on Urban Drainage (10ICUD)*, Copenhagen, Denmark, August 21-26 (2005).
28. Shumway, R.H., Azari, A.S. and Kayhanian, M. "Statistical approaches to estimating mean water quality concentrations with detection limits", *Environmental Science and Technology*, **36**(15), pp 3345-4453 (2002).
29. Kayhanian, M., Suverkropp, C., Ruby, R. and Tsay, K. "Multiple regression approach in predicting highway runoff pollutants concentration", *ASCE EWRI*, Salt Lake City, UT, June 27-July 1 (2004).