Environ Monit Assess (2012) 184:43–52
DOI 10.1007/s10661-011-1945-6

Optimization models for siting water quality monitoring stations in a catchment

Jehng-Jung Kao · Pei-Hao Li · Wen-Shin Hu

Received: 2 June 2010 / Accepted: 9 February 2011 / Published online: 8 March 2011

Abstract A water quality monitoring network (WQMN) must be designed so as to adequately protect the water quality in a catchment. Although a simulated annealing (SA) method was previously applied to design a WQMN, the SA method cannot ensure the solution it obtained is the global optimum. Therefore, two new linear optimization models are proposed in this study to minimize the deviation of the cost values expected to identify the possible pollution sources based on uniform cost (UC) and coverage elimination uniform cost (CEUC) schemes. The UC model determines the expected cost values by considering each subcatchment being covered by which station, while the CEUC model determines the coverage of each station by eliminating the area covered by any upstream station. The proposed models are applied to the Derchi reservoir catchment in Taiwan. Results show that the global optimal WQMN can be effectively determined by using the UC or CEUC model, for which both results are better than those from the SA method, especially when the number of stations becomes large.

Keywords Monitoring · Optimization · Site selection · Catchment · Water quality · Model

J.-J. Kao (✉) · W.-S. Hu
Institute of Environmental Engineering, National Chiao Tung University, 1001 University Road, Hsinchu 30010, Taiwan, Republic of China
e-mail: jjkao@mail.nctu.edu.tw

W.-S. Hu
e-mail: whhu.ev92g@nctu.edu.tw

P.-H. Li
Green Management Department, Industrial Technology Research Institute, Rm. 303A, Bldg. 64, 3F, 195, Sec. 4, Chung Hsing Rd., Chutung, Hsinchu 31040, Taiwan, Republic of China
e-mail: peihaoli@itri.org.tw

Introduction

Reservoirs are major sources of drinking water in Taiwan. However, human or agricultural activities in the upstream catchments usually contribute a significant amount of contaminants and thus deteriorate the water quality in the reservoirs. Establishing a proper sampling network is thus necessary to monitor and to ensure the water quality for protecting the health of the ecology as well as that of the human population.

As described by Dixon and Chiswell (1996), various issues, such as information goals, indicators, and data analyses, must be evaluated prior to designing an appropriate water quality monitoring network (WQMN). Dixon et al. (1999) indicated that the cost involved in finding the source of a pollution event should be considered when determining proper sampling sites. Furthermore, with a limited budget available, only a small number

of sampling sites can be installed in a large catchment. Therefore, it is a challenge to determine the appropriate locations for installing an effective WQMN that can provide representative water samples.

Several approaches have been proposed for the design of a monitoring network. For example, Sharp (1971) proposed a successive division algorithm to find topologically optimum sampling sites using the Shreve stream order number (Shreve 1967). This method may not be able to obtain the optimal placement of sampling sites (Dixon et al. 1999). Dixon et al. (1999) thus applied the simulated annealing (SA) method to design a WQMN with minimal total expected cost for identifying possible pollution sources. They proposed three geographical factors derived from the ratios of the number of reaches (REACH), upstream bank length (LENGTH), and sub-catchment areas (AREA) to represent the expected cost for locating the problem after a pollution event is detected at a sampling site. However, these three factors cannot directly represent the pollution distribution characteristics. Although other related researches (e.g., Icaga 2005; Ning and Chang 2005; Strobl et al. 2006, 2007, Karamouz et al. 2009a, b; Telci et al. 2009) considered the pollution distribution characteristics, they still did not consider the pollution detection capability of a WQMN for locating the source of a pollution event. Kao et al. (2008) thus proposed three additional cost factors based on estimated pollution potential, including total phosphorus (TP), total nitrogen (TN), and sediment (SED) loads, for improving the selection of sampling locations using the same SA method to locate possible pollution sources more efficiently. However, the SA method is a heuristic method that may not locate the true global optimum, or may result in a premature termination during a SA searching procedure. A similar situation may occur with other heuristic methods such as the genetic algorithm (Park et al. 2006; Ouyang et al. 2008; Karamouz et al. 2009a; Telci et al. 2009). Therefore, optimization models are desired.

Two new linear optimization models are thus proposed in this study based on uniform cost (UC) and coverage elimination uniform cost (CEUC) schemes to minimize the deviation of the distribution of a cost factor. The cost factor can be any of the three topographical cost factors proposed by Dixon et al. (1999) or the three pollution cost factors proposed by Kao et al. (2008). The UC model determines each sub-catchment to be covered by which station, while the CEUC model determines the coverage of a station by eliminating the area covered by each upstream station. The proposed models and the SA method adopted in previous studies were applied to the Derchi reservoir catchment in Taiwan. Results obtained from the proposed models and the SA method are compared and discussed.

Cost functions

Six cost factors are adopted to formulate the objective functions of the proposed optimization models. Each cost function is used separately as the single objective function in the proposed models for siting monitoring stations. The first three cost functions were adopted from Dixon et al. (1999) based on geographical topologies, and the other three pollutant distribution-based cost functions were proposed by Kao et al. (2008).

As defined by Dixon et al. (1999), the cost for investigating the possible sources of a detected pollution event can be defined by the following equation:

$$E_{\text{cost}} = \sum_i P_i \text{Ew}_i \tag{1}$$

where E_{cost} is the expected cost, P_i is the occurrence probability of a pollution event in sub-catchment i, and Ew_i is the expected effort required to locate the source once a pollution event is detected in sub-catchment i. The six cost functions are formulated as follows:

$$\sum_i \left(\frac{m_i}{m_0}\right) \log_2 m_i \tag{2}$$

$$\sum_i \left(\frac{L_i}{L_0}\right) \log_2 m_i \tag{3}$$

$$\sum_i a_i^2 \tag{4}$$

$$\sum_i p_i^2 \tag{5}$$

$$\sum_i n_i^2 \tag{6}$$

$$\sum_i s_i^2 \tag{7}$$

where m_i is the number of reaches in sub-catchment i, m_0 is the total number of reaches in the entire catchment, L_i is the bank length in sub-catchment i, L_0 is the total bank length in the entire catchment, a_i is the area of sub-catchment i, and p_i, n_i, and s_i are respectively the estimated TP, TN, and SED loads generated from sub-catchment i that are determined from AGNPS (Young et al. 1987) modeling simulations.

For cost function 2, the occurrence probability for potential pollution is defined to be proportional to the ratio of the number of reaches in a sub-catchment to the total number of reaches in the entire catchment. A binary search is applied to locate the pollution sources, and the mean number of samples required for the detection can be assumed as $\log_2 m_i$. A similar definition is applied to the cost factor of the total bank length formulated in Eq. 3. For cost functions 4 to 7, the occurrence probability and the expected cost required to locate pollution sources are both assumed to be proportional to the magnitude of the AREA, TP, TN, and SED loads of a catchment, respectively. For the pollution potential cost functions 5 to 7, pollution loads are primarily generated from distributed sources, and a high pollution load indicates the existence of a large area or a large number of pollution sources. A pollution event is likely to happen in a place with a high pollution load, and its probability is thus assumed to be proportional to its estimated load. Furthermore, the effort required to identify the source when an event occurs is also expected to be proportional to the estimated load. Thus, the cost functions are expressed by Eqs. 5 to 7.

Water quality monitoring network design models

The average effort required to locate the source of a pollution event can be minimized if the investigation costs of all monitoring stations are similar. Therefore, two linear programming models are proposed herein to determine the monitoring network with minimal deviation of investigation costs among stations.

Uniform cost model

In the UC model, each sub-catchment is covered by a specific station. The event investigation cost for each station is determined by summing the cost values of those sub-catchments which are covered by the station. The optimal WQMN can then be determined by minimizing the total cost deviation of the selected monitoring stations from the average cost value. The UC model is thus formulated as follows:

$$\text{Min} \quad \sum_{i=1}^{N} u_i + v_i \tag{8}$$

s.t.

$$x_{\text{ave}} - u_i + v_i - x_i = 0 \quad \forall i \tag{9}$$

$$x_{\text{ave}} = \frac{\sum x_i}{M} \tag{10}$$

$$y_1 = 1 \tag{11}$$

$$\sum y_i = M \tag{12}$$

$$z_{ij} \le 1 - y_k \qquad \forall j \in C_k, \forall k \in U_i, \forall i \tag{13}$$

$$\sum_{i \in s_j} z_{ij} = 1 \qquad \forall j \tag{14}$$

$$x_i < B\, y_i \qquad \forall i \tag{15}$$

$$x_i = \sum_{j \in C_i} z_{ij} F_j \qquad \forall i \tag{16}$$

$$y_i \in [0, 1] \tag{17}$$

$$0 \le z_{ij} \le 1 \qquad \forall i, j \tag{18}$$

where N is the number of reaches in the catchment; M is the number of stations to be installed; x_i is the cost value of candidate station i; x_{ave} is the average cost of all selected stations; u_i and v_i are the negative and positive deviations from the average cost, respectively; y_i is a 0–1 integer

variable, for which a value of 1 indicate station i is selected as a monitoring station; z_{ij} is a variable between 0 and 1, for which 1 indicates that reach j is covered by a candidate station i; C_k is the set of reaches covered by a candidate station k; U_i is the set of all the other candidate stations located in the upstream of station j; S_j is the set of candidate stations covering reach j; B is an extremely large value; and F_j is the cost value for investigating the pollution source in reach j.

The objective function minimizes the total cost deviation of selected monitoring stations from the average cost value. The most downstream station is always selected to be a monitoring station by Eq. 11 to ensure all reaches are covered. The total number of monitoring stations is restricted to a predefined number M in Eq. 12. The reaches covered by each station are determined according to the topology of the reaches. However, if there is already a selected station located in the upstream of a candidate station, then those reaches already covered by the upstream station are removed from the coverage of the candidate station, as constrained in Eq. 13. Equation 14 is applied to ensure that each reach in the catchment is covered by a monitoring station. The cost value of station i can then be evaluated by summing all the cost values of the covered reaches, as computed by Eq. 16. If station i is not selected, the cost value of station i will be set to be 0 by Eq. 17. Although z_{ij} is not a 0–1 integer variable, Eqs. 14 and 15 drive z_{ij} to be either 0 or 1. After all the cost values of selected stations are determined, the cost deviations of each station are then determined by Eq. 9, and the average cost value for all stations is computed by Eq. 10.

Coverage elimination uniform cost model

Although the UC model can obtain a proper monitoring network, a significant number of variables are required for determining the coverage of selected monitoring stations. The other model, called the CEUC model, determines that the coverage of stations by eliminating overlapped reaches is thus proposed to reduce the number of variables. The optimal WQMNs that are determined using both the CEUC and the UC models are expected to be the same, except in the case with multiple alternative optima. The CEUC model is formulated as follows:

$$\underset{s.t.}{\text{Min}} \quad \sum_{i=1}^{N} u_i + v_i + p_i L \tag{19}$$

$$x_{\text{ave}} - u_i + v_i - x_i = 0 \qquad \forall i \tag{20}$$

$$x_{\text{ave}} = \frac{\sum x_i}{M} \tag{21}$$

$$y_1 = 1 \tag{22}$$

$$\sum y_i = M \tag{23}$$

$$\sum x_i = X \tag{24}$$

$$x_i = A_i\, y_i - \sum_{j \in U_i} x_j + p_i \qquad \forall i \tag{25}$$

$$x_i < B\, y_i \tag{26}$$

$$B\,(1 - y_i) > p_i \qquad \forall i \tag{27}$$

$$y_i \in [0, 1] \tag{28}$$

where p_i is a dummy variable adopted to ensure x_i equal to 0 while station i is not selected as a monitoring station, L is a small value, A_i is the original investigation cost value of candidate station i, X is the total cost value of all reaches, and all the other variables are the same as those used in the UC model.

The objective function minimizes the summation of differences among the cost values of all selected stations. The term for the dummy variable p_i multiplied by a small value L is used to assure that the dummy variable is driven to zero, if at all possible.

The cost value of a selected monitoring station i is determined by Eq. 25. If station i is not selected, i.e., y_i is 0, x_i will be set to be 0 by Eq. 26 and subsequently $A_i y_i$ in Eq. 25 will also be 0. At the same time, variable p_i balances the cost values of the selected upstream stations and ensures that the cost value of each station is positive in Eq. 25. On the other hand, if station i is selected as a monitoring station, i.e., y_i is 1, then p_i is set to be 0 by Eq. 27 and can be ignored. The cost value

x_i can then be determined without considering the cost values of the stations upstream to station i. The total cost value of all selected stations is constrained to be equal to the total sum of the cost values in Eq. 24 to ensure that all reaches are covered by the determined monitoring network. The other equations are the same as those in the UC model.

Simulated annealing method

The SA method used by Dixon et al. (1999) and Kao et al. (2008) is also applied in this study for comparison with the proposed models. The SA method is briefly described as follows, and the detail of the method is referred to Dixon et al. (1999). The decision variables in the SA method are the locations for placing monitoring stations. A pre-defined initial temperature (C_0) is cooled down by a factor of (<1) in each SA iteration until the number of desired cooling steps or final temperature is reached. In every iteration, a monitoring station selected in the previous iteration is chosen randomly to make a random move to an upstream or a downstream reach. If the associated cost determined by a pre-specified cost function of the new WQMN is better than that in the previous iteration, a subsequent iteration is then initiated to continue the task. If a worse placement is generated, the iteration is continued while a generated random number, between 0 and 1, is less than a specified function, $\exp(-\Delta E/C)$, or the current placement is discarded and another placement is tried. The whole procedure is repeated until the specified number of temperature cooling steps has been achieved.

Case study

The study area used in our previous research (Kao et al. 2008) for the Derchi Reservoir, as shown in Fig. 1, is also used herein for comparison purposes. The Derchi Reservoir stores approximately 2.5×10^8 m^3 of water in central Taiwan and is a major source of local drinking water. The catchment area of the reservoir is about 602 km^2 and is divided into 63 sub-catchments. The land in the area is mainly used for orchards, with approximately 2,620 ha under cultivation. The TP

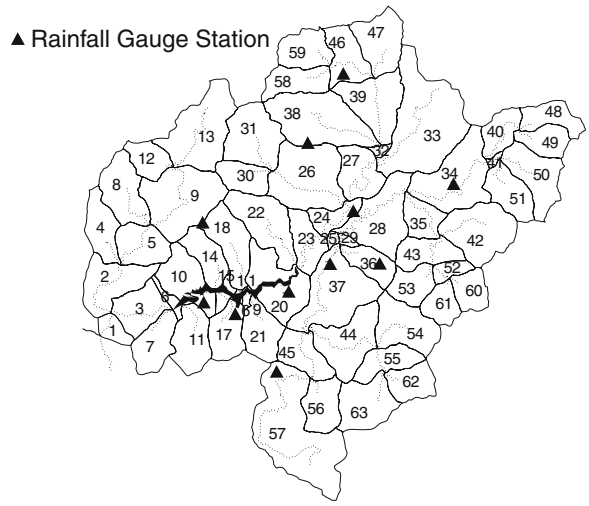

Fig. 1 The Derchi Reservoir in central Taiwan and the rainfall gauge stations

concentration of the reservoir water body ranges between 40 and 140 ppb. According to the index proposed by Carlson (1977), this reservoir is eutrophic, and water quality control is thus required.

The pollution distribution in the catchment had been estimated by the same modeling approach used by previous studies (Kao and Tsai 1997; Lin and Kao 2003; Kao et al. 2008) based on the simulation result provided by the AGNPS model (Young et al. 1987). The simulated NPSP distributions of TP, TN, and SED loads of sub-catchments are illustrated in Fig. 2. The sub-catchment with the higher NPSP load is marked in a darker color and vice versa. The distribution of the TP loads is similar to that of the TN loads, but the distribution of the SED loads is slightly different.

Results and discussion

The UC, CEUC, and SA models were formulated and applied to the Derchi Reservoir case. The number of monitoring stations was set to be between 2 and 20. For the stochastic characteristics of the SA model, the results obtained for the different runs may not be the same. The SA model was thus executed five times for each number of monitoring stations, and the best solution was selected for comparison with the results obtained by the other two models. The objective values of the

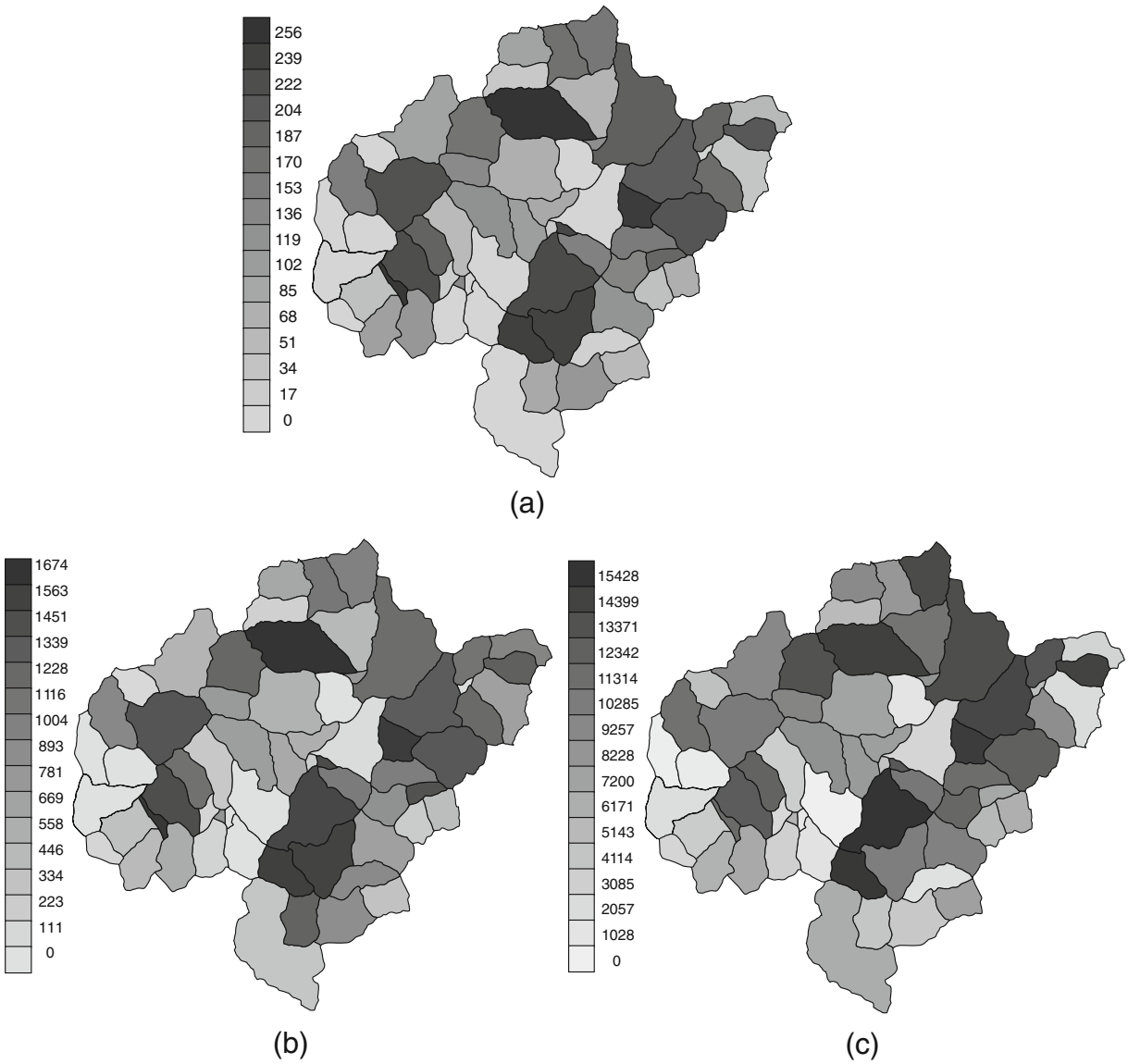

Fig. 2 Pollution loads in each sub-catchment: **a** TP, **b** TN, and **c** SED loads (in kilograms per day)

UC, CEUC, and SA models for varied numbers of monitoring stations are illustrated in Fig. 3. The WQMNs determined by the three models for eight monitoring stations, as shown in Fig. 4, were chosen for further analysis and comparison.

As shown in Fig. 3, the values for various cost functions decrease as the number of monitoring stations increases, i.e., the time required to trace the pollution source of a detected event is shortened. Both the UC and CEUC models minimize the differences in the cost values among selected monitoring stations, and the values determined by both models are almost the same. Although the SA approach also formulates the optimization model with the same objective as the UC and CEUC models, the WQMNs generated by the SA approach usually require more expected cost because the SA search may terminate prematurely before the global optimum is located and to locate the true global optimum by the SA search is generally time-consuming. Therefore, the cost function values of most WQMNs determined by the

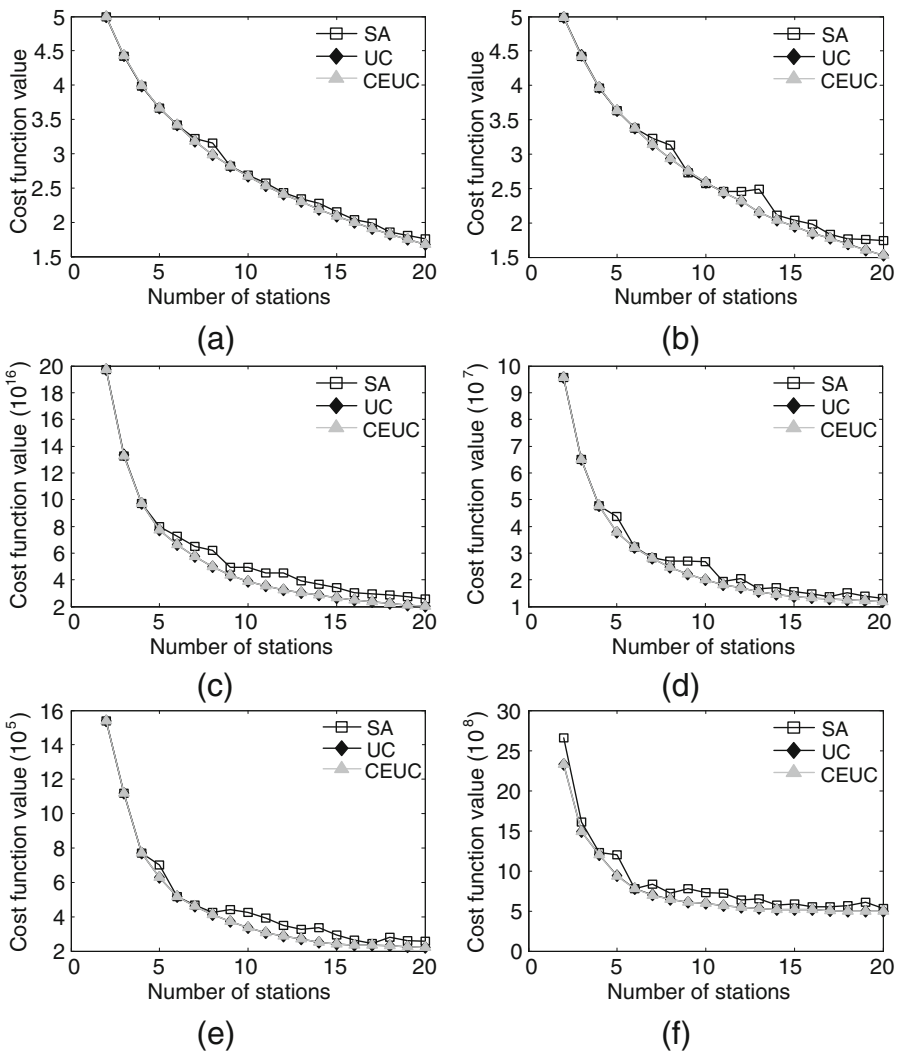

Fig. 3 Comparisons of the values obtained by the UC, CEUC, and SA models for the six cost factors: **a** REACH, **b** LENGTH, **c** AREA, **d** TN, **e** TP, and **f** SED

SA method are higher than those determined by the two proposed models, especially for those with a higher number of monitoring stations. Some marginally unreasonable results may be obtained from the SA method as well. For instance, as shown in Fig. 3f, the cost function value for the WQMN with nine monitoring stations is higher than that for eight monitoring stations. Since the solution space expands considerably when the number of monitoring stations increases, locating the global optimal WQMN by a SA search becomes quite difficult.

To evaluate the differences among the distributions of WQMNs obtained by the SA and those by the UC and CEUC models, the WQMNs with eight monitoring stations determined by the models are illustrated in Fig. 4. According to Fig. 4b, c, almost all monitoring stations selected by UC and by CEUC are the same, except for one station. The UC model selects sub-catchment 35, while sub-catchment 34 is selected by the CEUC model. However, the expected costs of the solutions obtained from both models are the same, and these two solutions are alternative optima. The WQMN determined by the SA method includes monitoring stations located near the branch ends in the catchment, such as sub-catchments 7, 22, and 57, as shown in Fig. 4a. The cost values of these monitoring stations are small with relatively small covered areas, and the distribution of the cost

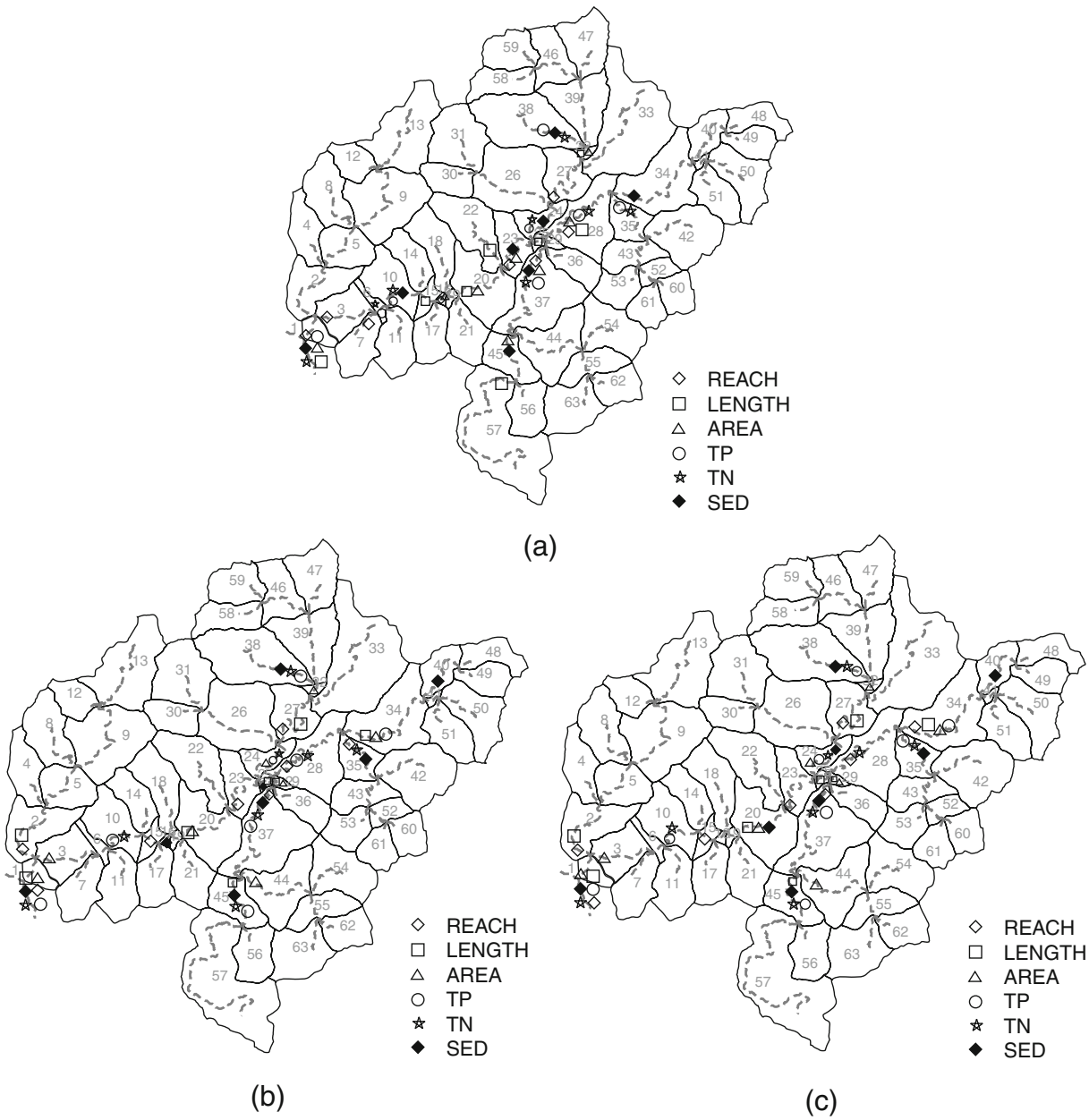

Fig. 4 Spatial distribution of the eight-station WQMNs determined based on different cost factors and different models: **a** SA, **b** UC, and **c** CEUC models

values among all monitoring stations will be uneven, leading to a higher cost function value. Therefore, as shown in Fig. 5, the cost function values obtained from the UC model are more evenly distributed than those determined by the SA method for both geography and pollution load-based cost factors. As the results illustrate in Fig. 3a, d, the even distribution of cost function values among the eight monitoring stations results in a low total expected cost.

When a small number of monitoring stations are established for the cost factors of area, TN, TP, and SED, the difference in cost function values among monitoring stations determined by

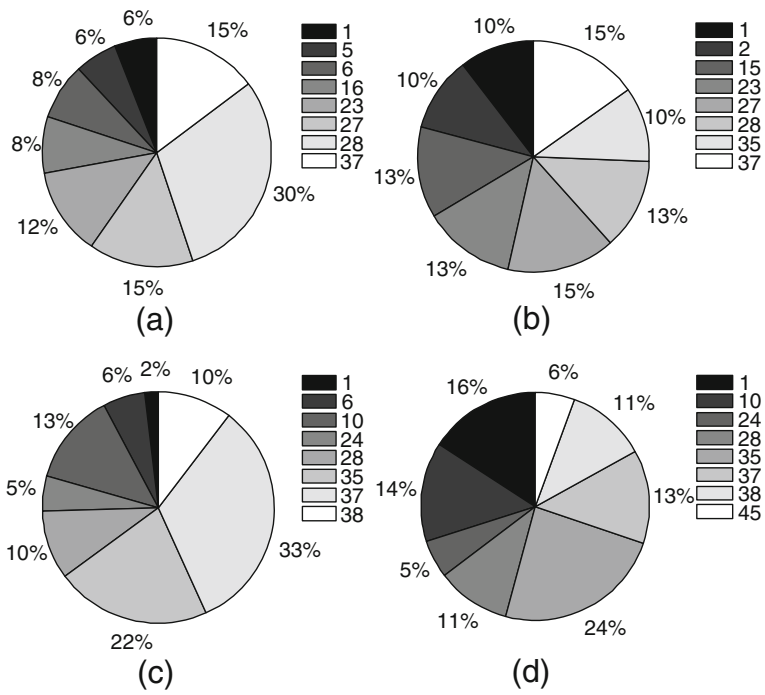

Fig. 5 The cost function value ratio for each monitoring station of the eight-station WQMNs determined by **a** the SA model with the REACH cost factor, **b** the UC model with the REACH cost factor, **c** the SA model with the TN cost factor, and **d** the UC model with the TN cost factor

all three methods becomes apparent; therefore, the expected costs are larger than those for a large number of monitoring stations, as shown in Fig. 3c–e. As the number of monitoring stations increases, those sub-catchments with significant high cost are evenly covered, and thus, the associated cost function values are reduced. Since the SA method may be prematurely terminated before the global optimum is found, almost all the cost function values of the WQMNs determined by the SA method are larger than those determined by the UC and CEUC models, especially for the WQMNs with a larger number of monitoring stations, although some of the differences are not significant.

About 40% to 65% of pollution loads of TP, TN, and SED are generated from five out of the total of 63 sub-catchments in the Derchi catchment, as shown in Fig. 2. Since these pollution loads are distributed unevenly in the catchment, the stations of the WQMNs determined are not evenly distributed neither. As illustrated in Fig. 4, sub-catchments 38 and 45 are selected by all three methods for the cost factors of TP, TN, and SED because they are effective to monitor the sub-catchments with high potential pollution loads. However, for the topographical cost factors of REACH, LENGTH, and AREA, the determined monitoring stations are evenly distributed, as shown in Fig. 4. Such an evenly distributed WQMN is not effective for monitoring the sub-catchments with high pollution loads that are likely to have pollution events occurring due to intensive human activities.

Conclusion

When the number of stations of a WQMN increases, the decision space will expand and a SA search may be prematurely terminated without locating the true global optimum. Thus, two optimization models, the UC and CEUC models, were proposed in this study to determine a proper WQMN with minimal cost to trace the source of a pollution event. Six topographical and pollution loading cost factors proposed by Dixon and Chiswell (1996) and Kao et al. (2008) were adopted in this study to formulate the optimization models. According to the results obtained of the case study for the Derchi catchment, although the SA method can determine WQMNs

with cost values close to those determined by the UC and CEUC models when the desired number of monitoring stations is small, the SA method can prematurely terminate with a poor WQMN if the parameters are not properly set. With the proposed UC and CEUC models, the true global optimal solution can be obtained. All the results obtained by the UC and CEUC models are consistent except for the cases with existing alternative optima.

The monitoring stations that were determined are distributed evenly among sub-catchments for the topographically based cost factors of REACH, LENGTH, and AREA. However, since the intensive agricultural activities are usually located diversely among the sub-catchments, the possible pollution sources with a considerable impact upon the water quality in the catchment are thus distributed unevenly among the sub-catchments. The WQMNs determined for the topographical factors might thus not be able to detect the pollution events efficiently. Therefore, the three pollution-based factors of TN, TP, and SED proposed by Kao et al. (2008) should be taken into consideration while determining a proper WQMN to protect the water quality in a more efficient manner. Different combinations of monitoring stations may be obtained if different cost factors are considered. A proper WQMN should be determined based on the possible pollution pattern of the study area and the strategy adopted by the authority to implement a monitoring program.

Acknowledgements The authors would like to thank National Science Council, Republic of China for providing partial financial support of this research under Grant NO. 96-2221-E-009-056-MY3 and 90-2211-E-009-019.

References

Carlson, R. E. (1977). A trophic state index for lakes. *Limnology and Oceanography, 22*, 361–369.

Dixon, W., & Chiswell, B. (1996). Review of aquatic monitoring program design. *Water Research, 30*(9), 1935–1948.

Dixon, W., Smyth, G. K., & Chiswell, B. (1999). Optimized selection of river sampling sites. *Water Research, 33*(4), 971–978.

Icaga, Y. (2005). Genetic algorithm usage in water quality monitoring networks optimization in Gediz (Turkey) river basin. *Environmental Monitoring and Assessment, 108*, 261–277.

Kao, J.-J., & Tsai, C.-H. (1997). Multiobjective zone TP reduction analyses for an off-stream reservoir. *Journal of Water Resources Planning and Management, 123*(4), 208–215.

Kao, J.-J., Li, P.-H., & Lin, C.-L. (2008). Siting analyses for water quality sampling in a catchment. *Environmental Monitoring and Assessment, 139*, 205–215.

Karamouz, M., Kerachian, R., Akhbari, M., & Hafez, B. (2009a). Design of river water quality monitoring networks: A case study. *Environmental Modeling and Assessment, 14*, 705–714.

Karamouz, M., Nokhandan, A. K., Kerachian, R., & Maksimovic, C. (2009b). Design of on-line river water quality monitoring systems using the entropy theory: A case study. *Environmental Monitoring and Assessment, 155*, 63–81.

Lin, Y.-C., & Kao, J.-J. (2003). Effects of seasonal variation in precipitation on estimation of non-point source pollution. *Water Science and Technology, 47*(7–8), 299–304.

Ning, S. K., & Chang, N. B. (2005). Screen the relocation strategies of water quality monitoring stations by compromise programming. *Journal of the American Water Resources Association, 41*(5), 1039–1052.

Ouyang, H.-T., Yu, H., Lu, C.-H., & Luo, Y.-H. (2008). Design optimization of river sampling network using genetic algorithms. *Journal of Water Resources Planning and Management, 134*(1), 83–87.

Park, S.-Y., Choi, J. H., Wang, S., & Park, S. S. (2006). Design of a water quality monitoring network in a larger river system using the genetic algorithm. *Ecological Modelling, 199*, 289–297.

Sharp, W. E. (1971). A topologically optimum water-sampling plan for rivers and streams. *Water Resources Research, 7*(6), 1641–1646.

Shreve, R. L. (1967). Infinite topologically random channel networks. *Journal of Geology, 75*, 178–186.

Strobl, R. O., Robillard, P. D., Shannon, R. D., Day, R. L., & Mcdonnell, A. J. (2006). A water quality monitoring network design methodology for the selection of critical sampling points: Part I. *Environmental Monitoring and Assessment, 112*, 137–158.

Strobl, R. O., Robillard, P. D., & Debels, P. (2007). Critical sampling points methodology: Case studies of geographically diverse watersheds. *Environmental Monitoring and Assessment, 129*, 115–131.

Telci, I. T., Nam, K., Guan, J., & Aral, M. M. (2009). Optimal water quality monitoring network design for river systems. *Journal of Environmental Management, 90*(10), 2987–2998.

Young, R. A., Onstad, C. A., Bosch, D. D., & Anderson, W. P. (1987). *AGNPS, agricultural non-point source pollution model.* Minnesota: USDA-ARS.