Journal of Environmental Monitoring

Dynamic Article Links

Cite this: *J. Environ. Monit.*, 2011, **13**, 2886

www.rsc.org/jem

PAPER

Spatiotemporal classification of environmental monitoring data in the Yeongsan River basin, Korea, using self-organizing maps

Y.-H. Jin,*[a] A. Kawamura,[b] S.-C. Park,[c] N. Nakagawa,[d] H. Amaguchi[e] and J. Olsson[f]

Received 9th February 2011, Accepted 10th August 2011
DOI: 10.1039/c1em10132c

Environmental monitoring data for planning, implementing and evaluating the Total Maximum Daily Loads (TMDL) management system have been measured at about 8-day intervals in a number of rivers in Korea since 2004. In the present study, water quality parameters such as Suspended Solids (SS), Biochemical Oxygen Demand (BOD), Dissolved Oxygen (DO), Total Nitrogen (TN), and Total Phosphorus (TP) and the corresponding runoff were collected from six stations in the Yeongsan River basin for six years and transformed into monthly mean values. With the primary objective to understand spatiotemporal characteristics of the data, a methodologically systematic application of a Self-Organizing Map (SOM) was made. The SOM application classified the environmental monitoring data into nine clusters showing exclusively distinguishable patterns. Data frequency at each station on a monthly basis identified the spatiotemporal distribution for the first time in the study area. Consequently, the SOM application provided useful information that the sub-basin containing a metropolitan city is associated with deteriorating water quality and should be monitored and managed carefully during spring and summer for water quality improvement in the river basin.

Introduction

Anthropogenic activities in river basins such as intensive land use and development with rapid growth of population have increased the amount of pollutants discharged into rivers worldwide, resulting in a substantial deterioration of water quality and degradation of river environments.[1,2] Such intensive development has also been carried out in a number of river basins in Korea and its environmental impact has negatively influenced the water quality in the rivers.

Therefore, the Ministry of Environment (ME) in Korea has been trying to prevent rivers and the surrounding environment from any kind of pollution by launching the Comprehensive Water Management Measures in 1996 and the Comprehensive Measures of the Four Main Rivers (*i.e.*, the Han River, the Nakdong River, the Geum River and the Yeongsan River) in 1998.[3] In particular, the recent focus on the environment in rivers

Environmental impact

Environmental monitoring data for planning, implementing and evaluating the Total Maximum Daily Loads (TMDL) management system have been measured at about 8-day interval and accumulated in a number of rivers in Korea since 2004. In the present study, a Self-Organizing Map (SOM) was systematically applied to classify the data including water quality parameters and the corresponding runoff measured from a river basin in Korea for the first time with the primary objective to understand spatiotemporal characteristics of the data. To conclude, the SOM application is substantially useful to examine the spatiotemporal distribution of data and provided meaningful information for comprehensive improvement of water quality condition in the study area. Visible representation of the spatiotemporal distribution was displayed by the spatiotemporal mesh proposed in this study.

[a] *Department of Civil and Environmental Engineering, Tokyo Metropolitan University, Japan. E-mail: nmdrjin@gmail.com; Fax: +81-42-677-2772; Tel: +81-42-677-2787*

[b] *Department of Civil and Environmental Engineering, Tokyo Metropolitan University, Japan. E-mail: kawamura@tmu.ac.jp; Fax: +81-42-677-2772; Tel: +81-42-677-2787*

[c] *Department of Civil Engineering, Dongshin University, Korea. E-mail: psc@dsu.ac.kr; Fax: +82-61-330-3161; Tel: +82-61-330-3135*

[d] *Department of Civil and Environmental Engineering, Tokyo Metropolitan University, Japan. E-mail: nakanaok@tmu.ac.jp; Fax: +81-42-677-2772; Tel: +81-42-677-2787*

[e] *Department of Civil and Environmental Engineering, Tokyo Metropolitan University, Japan. E-mail: amaguchi@tmu.ac.jp; Fax: +81-42-677-2772; Tel: +81-42-677-2779*

[f] *Research and Development (Hydrology), Swedish Meteorological and Hydrological Institute, Sweden. E-mail: jonas.olsson@smhi.se; Fax: +46-11-495-8001; Tel: +46-11-495-8322*

emphasizes that nonpoint source pollution should be more strictly controlled than before for watershed management and water quality improvement, because of its increasing relative impact compared to point source pollution.[4]

Furthermore, the Total Maximum Daily Loads (TMDL) management system in Korea has been implemented to achieve the water quality targets within an intended timeframe since it was initially introduced in the Special Comprehensive Measures on the Han River watershed in 1998.[3] The TMDL consists of three main procedures: a basic plan, an implementation plan and an implementation review. Local governments should establish implementation plans for the TMDL by compulsion if they fail to meet the water quality targets allocated by the basic plans of the TMDL. Implementation reviews of the local governments are conducted to evaluate whether or not the implementation plans are carried out properly.[3]

The ME in Korea has also established Water Environment Research Centers for the four main rivers. At these centers, water quality and corresponding runoff data have been measured at about 8-day intervals since 2004 in the outlets of numerous sub-basins in the four major river basins.[5] Various data analysis tools may be developed and applied to the observations to detect key features and improve the understanding of the governing processes, with the objective of more efficient watershed management and, in turn, improving water quality. However, the environmental monitoring data were rarely analyzed and utilized, although they could play an important role in planning new environmental policies as well as implementing and evaluating the TMDL. A few researchers have examined spatiotemporal characteristics of water quality in the Han River and the Nakdong River basins.[6,7] However, no simultaneous spatiotemporal characterization of water quantity and water quality has yet been performed in Korea.

The four main river basins in Korea generally contain one or two large intensively developed and densely populated cities. The sub-basins including cities have negatively influenced the water quality and overall environment of the rivers with larger amounts of pollutants than in other sub-basins. Hence, it is crucial to understand, evaluate and compare the environmental impact of each sub-basin in order to efficiently design environmental policies and measures for sustainable improvement of water quality and watershed management.

However, it is generally difficult to understand the spatiotemporal characteristics of multivariate data such as the environmental monitoring data in a river basin as a whole. It can be meaningful to obtain and extract information from the classification of multivariate data into a number of patterns representing different characteristics. In other words, it is considerably applicable and feasible to divide the overall environmental aspect into a number of characteristic patterns. Recently, the Self-Organizing Map (SOM) has been frequently used as a powerful and effective data analysis tool for detection of data characteristics by pattern classification and visualization onto two-dimensional arrays. The application of SOM has been reported in diverse research fields such as ecology,[8–14] geomorphology,[15,16] hydrology,[17–23] meteorology,[24,25] and wastewater treatment.[26,27] Previous work by the authors has demonstrated the efficiency of SOMs to evaluate vulnerability to erosion[16] and to classify complex nonlinear synoptic fields for identifying heavy rainfall events.[25]

In terms of water quality applications, the SOM has been used for analyses of coastal water quality, sediment contamination and detection of abnormal water quality changes.[28–30] Other applications include the assessment of various impact sources on a river and the evaluation of spatiotemporal patterns.[31,32] In conjunction with a statistical analysis and a decision support system, SOMs were utilized for classification, modeling, interpretation and assessment of the water quality in rivers.[33,34] The above studies applied SOMs for the respective purposes without taking into account the water quantity such as river runoff. However, when water quality parameters are analyzed by a SOM, it is crucial to include also the corresponding runoff data in the analysis because the concentration data are significantly influenced by runoff.

As mentioned above, the applicability of SOMs has been demonstrated in a number of studies using a variety of data types from diverse research fields. However, although the SOM structure is extremely sensitive to the data transformation method, most studies perform data transformation and SOM structure determination without considering the close relationship between them. The distribution of raw data should be initially investigated to select a proper data transformation method which can guarantee the appropriate application of SOMs. Thus, the application of SOMs must be carried out in a methodologically systematic way and not as a black box method. To date, no such systematic application of SOMs has been carried out for environmental monitoring data from river basins in Korea.

Therefore, in the present study, pattern classification analysis by SOMs combined with a hierarchical cluster analysis is carried out with the primary objective of better understanding the governing processes by a detailed spatiotemporal characterization of the environmental monitoring data. The research utilizes the data measured in the Yeongsan River basin, which have never been used before for this purpose. Water quality concentrations and the corresponding runoff data are jointly examined for simultaneous consideration of both water quality and quantity. A methodologically systematic approach for SOM application with appropriate methods for data transformation, map size determination and SOM training is applied in order to ensure robust and credible results.

Study area and data used

For the implementation and evaluation of the TMDL in Korea, environmental monitoring data with the corresponding runoff have been measured in the outlets of sub-basins at about 8-day intervals since 2004. The Yeongsan River basin, which is one of the four main rivers in Korea and located in the southwestern part of Korean peninsula, was chosen for the present study because the data have not yet been much analyzed.

The river basin has an area of 3455 km^2 and the length of the main stream is 136 km. Environmental monitoring data from six stations in the river basin are available for the study as shown in Fig. 1. Two of the six stations are located in tributaries and are named T1 and T2, respectively, while the rest of the stations (M1, M2, M3 and M4) are located in the main stream from upstream to downstream. In the center of the river basin, the Gwangju metropolitan city is situated with a population of more than

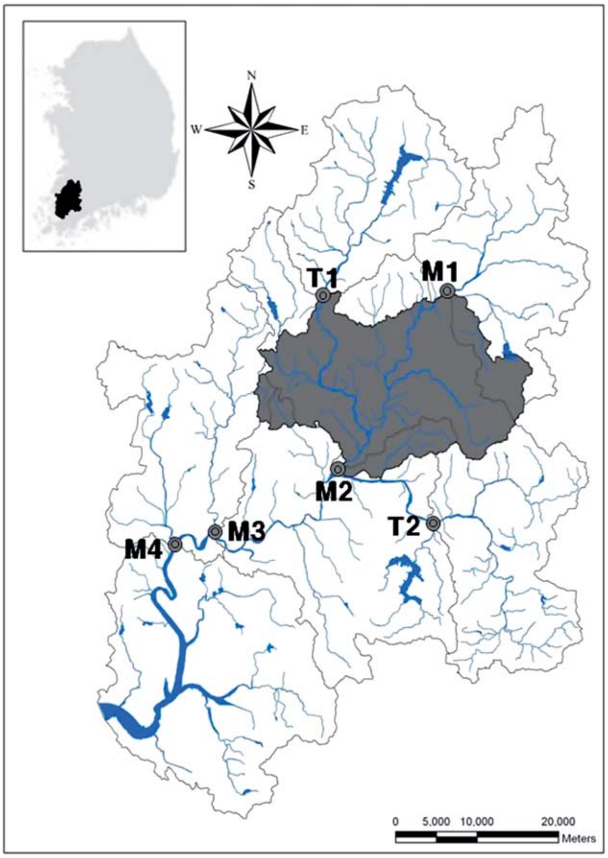

Fig. 1 Location map of the six stations (T1, T2, M1, M2, M3 and M4) in the Yeongsan River basin situated in the southwestern region of Korea, showing the Gwangju metropolitan city with gray area.

1.4 million. The city has a significant influence on the water quality of the main stream, which has been measured in the M2 station.[35]

The environmental monitoring data include runoff and the water quality concentration parameters such as Suspended Solids (SS), Biochemical Oxygen Demand (BOD), Dissolved Oxygen (DO), Total Nitrogen (TN), and Total Phosphorus (TP). The data have been measured in the six stations at about 8-day intervals containing missing values in some months. Therefore, the raw data were transformed into monthly mean values during the 6-year data period from September 2004 to August 2010.

The monthly mean values must be transformed properly before the application of SOM so that all parameters are given the same or similar importance.[12] In particular, the results of the SOM application are highly sensitive to the data pre-processing method used, because the SOM is trained so as to be organized according to the Euclidean distances between input data.[29] Three methods for data pre-processing such as log-transformation,[11,12,29] range scaling into [0, 1][12,25,29] and variance scaling by mean values and standard deviations of respective parameters[26,29] are generally used for standardization or normalization of the data used.

In the present study, the skewness of each parameter's frequency distribution was initially analyzed by plotting histograms as shown in Fig. 2(a) with BOD as an example. Log-transformation was applied to reduce the positive skewness (*e.g.*, Fig. 2(b)) for all parameters except for DO which did not have any clear skewness. In addition, the log transformation can smooth the data and reduce the influence of extreme values.[12] Without the application of log transformation, the biased distribution may remain causing inappropriate classification by SOM. Then, variance scaling was carried out for all parameters so that the transformed data were distributed symmetrically with the same mean value and standard deviation as shown in Fig. 2(c).

Configuration of SOM

SOMs, a kind of unsupervised Artificial Neural Network (ANN), can project high-dimensional information onto a low-dimensional format, usually a two-dimensional hexagonal array. It provides a readily understandable and visualized result of pattern classification. However, in addition to the visualization, the eventual purpose of the SOM application is to obtain useful and informative reference vectors. The reference vectors are also known as weights, connection weights, prototype weights,

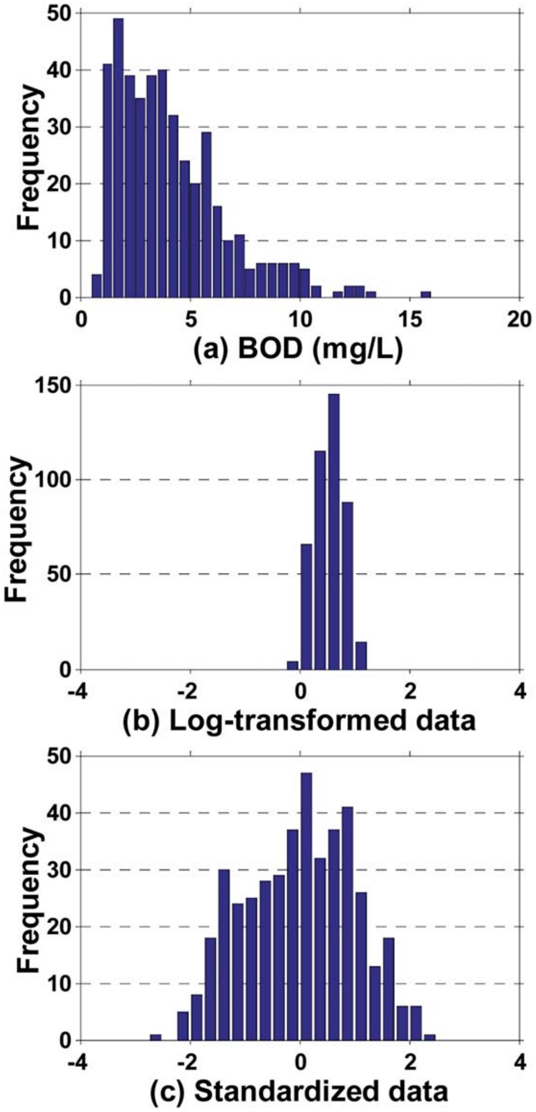

Fig. 2 Histograms for (a) raw data, (b) log-transformed data and (c) standardized data of BOD.

generalized median and codebook.[12,17,22,24] The vectors can be obtained after iterative updates through a training phase consisting of three main procedures: competition between nodes, selection of a winner node and update of the reference vectors.[25,36]

There are several key issues such as determination of the total number of nodes (*i.e.*, map size) and side lengths for constructing the SOM and selection of a proper initialization method as well as a data transformation method, for ensuring that the purpose of the SOM application is achieved. It is often useful to apply other clustering approaches for pattern classification using the reference vectors as a fine-tuning phase after the SOM training, in addition to the *U*-matrix (unified distance matrix) which is commonly used to create a rough visualization of the classification.

In order to determine the SOM structure, a heuristic rule of $m = 5\sqrt{n}$ is generally used, with m denoting the total number of nodes and n for the number of input data. In general, the larger the map size that is determined, the more detailed patterns can be identified. However, the topographical proximity of clusters is reduced. The heuristic rule can offer an optimized map size simultaneously considering the accuracy of pattern classification and topographical adjacency among clusters.[12,26] As the rule expresses that the total number of nodes varies with respect to the number of data used, a suitable time scale of the data should be determined prior to the application of the SOM.

The ratio of side lengths for the SOM is determined by the ratio between the two largest eigenvalues of the input data.[16,26,29,36] It should be noted that the ratio is strongly dependent on the data transformation method because the two maximum eigenvalues are highly dependent on how the data are transformed. Therefore, the application of an appropriate transformation method for the data is critical for obtaining an appropriate SOM structure.

After the determination of the SOM structure, each node is set with a reference vector by an initialization method. Recently, the linear initialization method is preferably used because it can improve the training phase.[14] Further, when only limited data are available, the linear initialization is more suitable for the pattern classification than the random initialization, as the latter requires a large dataset and might cause boundary effects near the edges of the map.[14,16,25] In addition, the linear initialization can use eigenvalues and eigenvectors of the input data to set the initial reference vectors on the structured SOM. It means that the initial reference vectors already include prior information about the input data, resulting in an acceleration of the training phase.[36] Iterative updates of the reference vectors are carried out by a training algorithm for which the batch mode is usually used.[26,29] The reference vectors obtained at the end of the training phase can be fine-tuned using cluster analysis methods.

Two main categories of cluster algorithms have been applied for fine-tuning of reference vectors.[26] One of them is known as the partitional clustering algorithm, in which the most frequently used method for SOMs is the *k*-means algorithm. The optimal number of clusters is selected by the Davies–Bouldin Index (DBI) which is calculated based on similarity within a cluster and dissimilarity between clusters. The DBI values are calculated from a minimum of two clusters to the total number of nodes. The number of clusters showing the minimum DBI is the optimal for the trained SOM.[12,24–26] The second main category is hierarchical cluster analyses, in which Ward's linkage method is the most commonly applied.[9,10,13,16]

In the present study, the SOM structure is determined by the heuristic rule for the total number of nodes and the side lengths are determined by the ratio between the two maximum eigenvalues of the transformed data. Reference vectors for the SOM with the commonly used hexagonal array are initially set using the linear initialization method to improve the training phase by the batch mode, as mentioned above. The optimal number of clusters is determined by the minimum DBI using *k*-means algorithm and a final fine-tuning cluster analysis is carried out by Ward's method with the optimal number of clusters.

Pattern classification

Based on the methodologically systematic configuration described above, a SOM size of 96 nodes (a hexagonal array with 16 nodes for a vertical direction and 6 for a horizontal direction) was used for pattern classification of the standardized environmental monitoring data. Fig. 3 shows the obtained component planes of the reference vectors of all six parameters, which were standardized into [0, 1]. Comparison between the component planes can indicate informative and qualitative relationships between parameters of concern.[16,29] For example, the component planes of runoff (Fig. 3(a)) and SS (Fig. 3(b)) reveal that the two parameters have a strong correlation as seen by the similar increase in shade from the upper right part to the lower left. The component planes of BOD, TN and TP are also strongly positively correlated (Fig. 3(c), (e) and (f)); however, no clear correlation with any other parameter is emergent for DO (Fig. 3(d)).

Table 1 quantitatively confirms the strength of relationship between parameters using the standardized reference vectors. The highest correlation coefficient of 0.98 was shown between TN and TP. The relationship of BOD with SS, TN and TP indicated significantly high correlation coefficients of 0.70, 0.88 and 0.93, respectively. DO mainly revealed inversed correlations with other parameters except for TN but the correlation coefficients were relatively low as shown in the table.

In order to select the optimal number of clusters for the configured SOM, the DBI values based on the *k*-means clustering algorithm were calculated for the possible minimum number of clusters (2) to the maximum number (96). Fig. 4 represents the variation of DBI values and its front part between two and twenty clusters was magnified with a logarithmic scale for the vertical axis to show the minimum DBI visibly. The minimum DBI is found for nine clusters, which is thus the most appropriate number for the pattern classification of the environmental monitoring data. Subsequently, the hierarchical clustering algorithm by Ward's method was applied for the nine clusters to fine-tune the pattern classification.

Fig. 5 shows the hierarchical cluster tree, which is also known as a dendrogram, with the nodes of the SOM classified into nine different clusters. Fig. 6 shows the pattern classification map of the nine clusters in which the numbers of data classified into each node are also given. Simultaneous consideration of the component planes (Fig. 3) and the pattern classification result (Fig. 6) indicates what kind of data the respective clusters

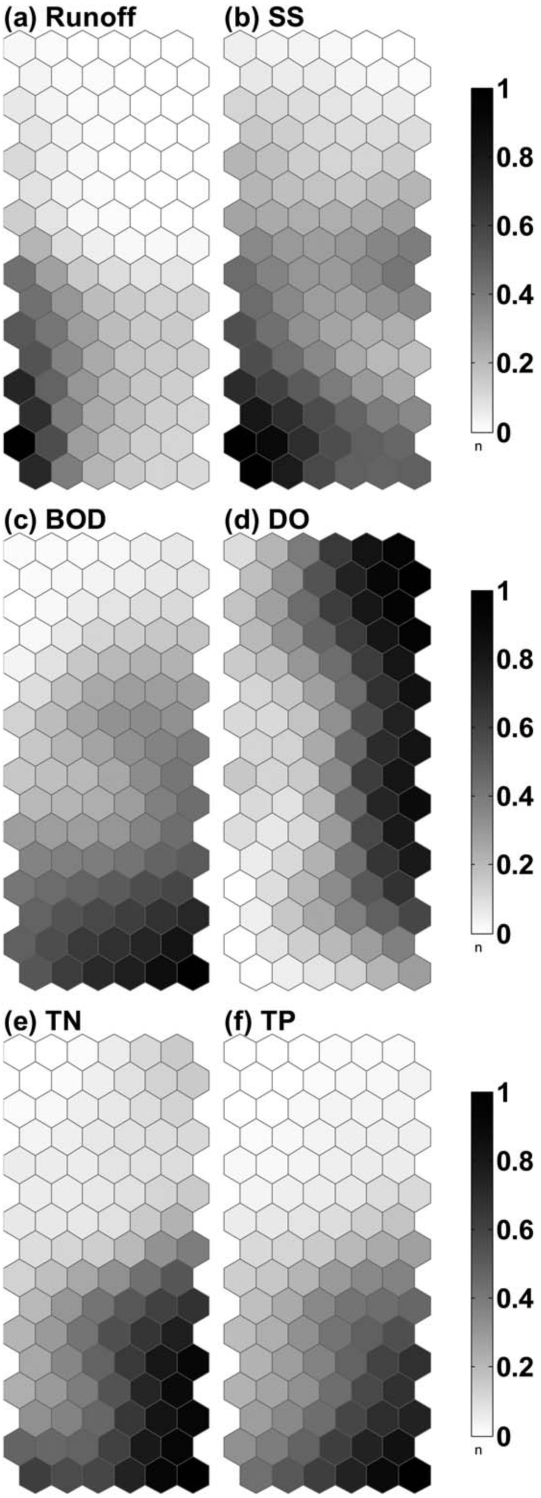

Fig. 3 Component planes for (a) runoff, (b) SS, (c) BOD, (d) DO, (e) TN and (f) TP.

include. On the one hand, cluster-1 situated in the upper left part of Fig. 6 is associated with high water quality characterized by low runoff, SS, BOD, TN, TP and relatively high DO. This pattern is seen in the same part of the respective component planes for each parameter as shown in Fig. 3. On the other hand, the worst water quality condition with extremely high BOD, TN and TP, significantly high SS, low DO and relatively low runoff located in the lower right part of each component plane as shown in Fig. 3 is associated with cluster-5 shown in Fig. 6.

Table 1 Correlation coefficients between reference vectors for each parameter

	SS	BOD	DO	TN	TP
Runoff	0.86	0.38	−0.59	0.30	0.30
SS		0.70	−0.55	0.51	0.56
BOD			−0.16	0.88	0.93
DO				0.03	−0.09
TN					0.98

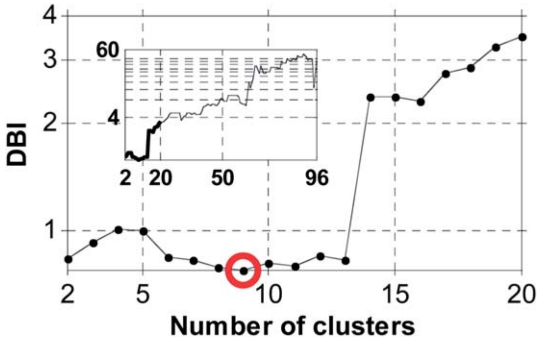

Fig. 4 Variation of DBI values with the optimal number of clusters marked by the circle on the figure.

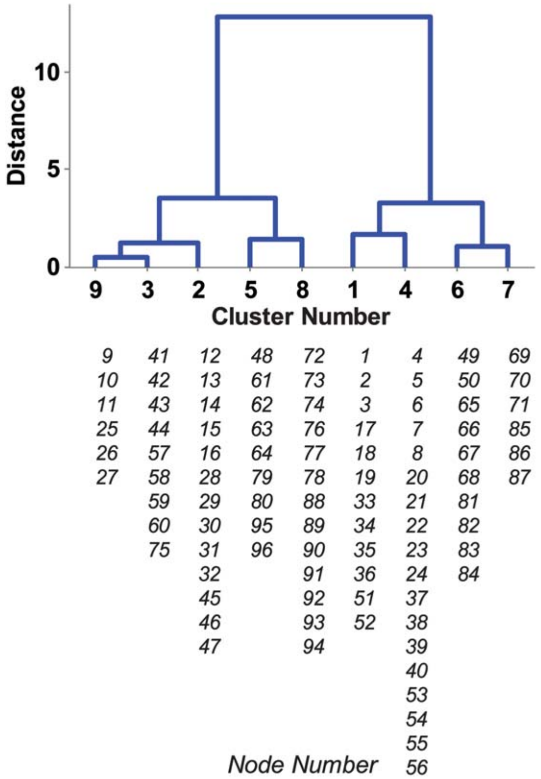

Fig. 5 Dendrogram with node numbers classified into the respective clusters.

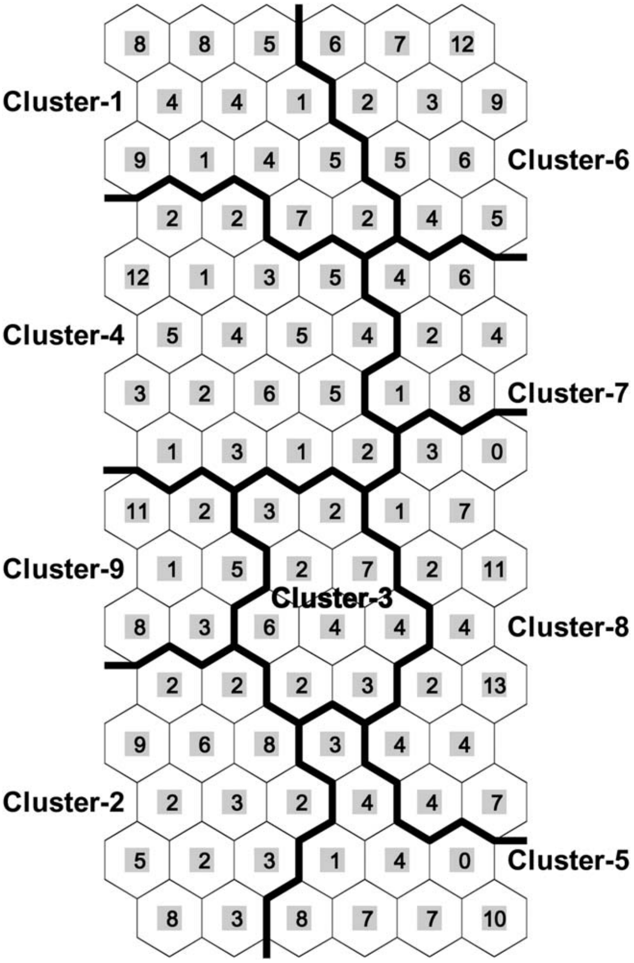

Fig. 6 Pattern classification map of the nine clusters by the SOM. The numbers on squares of the map represent the number of data classified into each node.

Further, more quantitative information than the visualized pattern classification can be extracted and interpreted from the obtained reference vectors. The first quartile, median (*i.e.*, the second quartile) and the third quartile for the respective clusters were calculated using the standardized reference vectors in order to numerically characterize the classified data. For example, the quartiles for cluster-1 were calculated using the standardized reference vectors of the 12 nodes classified into the cluster.

Fig. 7 displays radar charts of the six parameters for the nine clusters with the first quartile, median and third quartile plotted. The most ideal water quality condition can be defined as value 0 for SS, BOD, TN, TP and 1 for DO in Fig. 7.

The visible patterns of cluster-6 (Fig. 7(f)) and cluster-7 (Fig. 7 (g)) are similar as shown in the figure. The pattern with the highest DO, significantly low SS, BOD, TN and TP is associated with cluster-6 representing the best water quality condition of all clusters. Cluster-7 represents a similar water quality condition as cluster-6 but with slightly higher SS, BOD, TN, TP and lower DO. The lowest values of SS, BOD, TN, and TP with slightly low DO are classified into cluster-1 (Fig. 7(a)). Cluster-4 shown in Fig. 7(d) is characterized by significantly low SS, BOD, TN, TP and low DO.

The patterns of cluster-5 and cluster-8 are relatively similar. The worst water quality condition is represented by cluster-5 (Fig. 7(e)) in which the highest BOD, TN, TP with significantly high SS and relatively low DO were classified. Cluster-8 (Fig. 7 (h)) includes relatively high SS, BOD and significantly high TN, TP, DO. Cluster-3 (Fig. 7(c)) and cluster-9 (Fig. 7(i)) commonly include relatively high SS, BOD, TN, TP and low DO. However, cluster-3 is associated with higher BOD, DO, TN and TP values and cluster-9 is associated with higher runoff and SS values. The pattern including the highest runoff, SS and the lowest DO was classified into cluster-2 (Fig. 7(b)). Cluster-2 and cluster-9 include higher runoff values than the other clusters.

The classified nine clusters could be divided into two main environmental patterns. Relatively better water quality conditions were associated with cluster-6, 7, 1 and 4 shown in the right part of Fig. 5 and in the upper part of Fig. 6. The other group included cluster-3, 9, 8, 2 and 5, representing relatively worse water quality conditions as seen in the left part of Fig. 5 corresponding to the lower part of Fig. 6.

Table 2 shows the mean values calculated from raw data of each parameter for whole data and the classified data into the respective clusters. The runoff for cluster-6, 7, 1 and 4 indicates much lower mean values than the whole data. Cluster-6 and 7 represented lower mean values for the pollutants such as SS, BOD, TN and TP with higher DO concentration than those for the whole data. It confirms that the two clusters included the high water quality as mentioned above. Cluster-1 and 4 showed slightly higher mean values for runoff and significantly lower mean values for DO than those of cluster-6 and 7. The mean values for the pollutants in cluster-1 and 4 were lower than those for the whole data.

However, cluster-3, 9, 8, 2 and 5 showed the range between slightly lower and much higher mean values for runoff comparing to the mean value for the whole data. Considering water quality parameters, the clusters showed the general pattern with high mean values for pollutants and low mean values for DO. In particular, cluster-8, 2 and 5 represented considerably higher mean values for the pollutants than those for the whole data, showing seriously deteriorated water quality.

In addition, the frequency of data classified into each cluster was investigated in the respective stations on a monthly basis for better understanding of the spatiotemporal variability. Fig. 8 and 9 display spatiotemporal meshes for the two main groups mentioned above with the number of data occurrences counted. The horizontal axis in the mesh represents each month while the vertical axis shows the six stations from upstream to downstream. The thick dashed horizontal lines in the meshes represent the conceptual location of the Gwangju metropolitan city situated between M1 and M2 stations. The maximum data frequency of a particular month and station is 6 because the data measurement period is six years. The sums of data frequencies for each station are shown in the column to the right of the meshes.

On the one hand, from Fig. 8, only the environmental monitoring data in the three upstream stations T1, T2 and M1 (Fig. 1) were classified into the clusters showing relatively better water quality conditions. On the other hand, from Fig. 9, the data associated with relatively worse water quality conditions were mostly from the three downstream stations M2, M3 and M4. In particular, station M2 shows the highest frequency in cluster-5

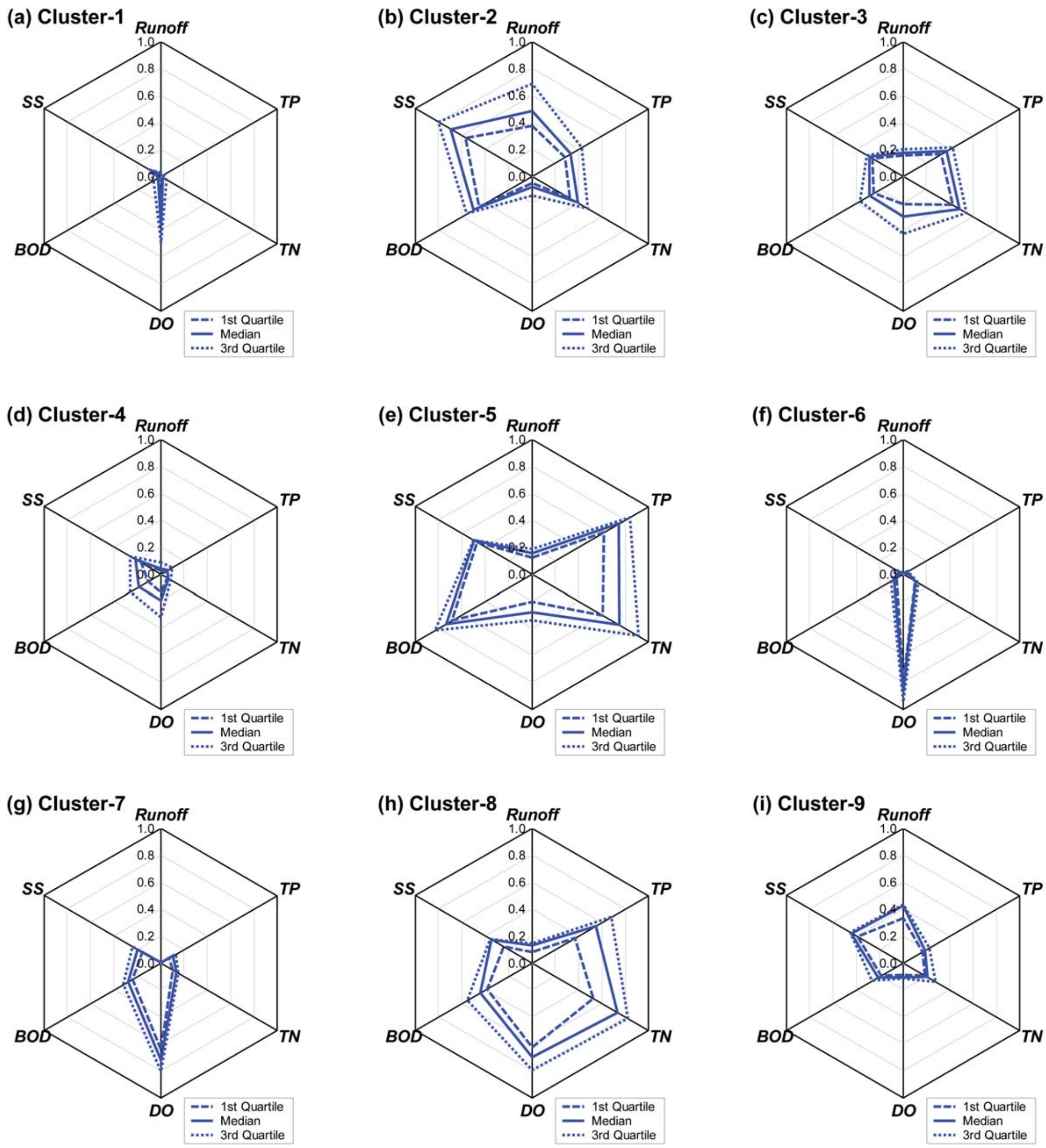

Fig. 7 Radar charts for the respective clusters with the first quartile (dashed lines), median (solid lines), and the third quartile (dotted lines) by the obtained reference vectors.

Table 2 Mean values calculated from raw data of each parameter for whole data and the classified data into the respective clusters

Mean	Runoff/m^3 s^{-1}	SS/mg L^{-1}	BOD/mg L^{-1}	DO/mg L^{-1}	TN/mg L^{-1}	TP/mg L^{-1}
Whole data	31.07	17.27	4.16	10.00	4.72	0.29
Cluster-1	7.39	6.79	1.70	9.26	1.95	0.09
Cluster-2	121.31	38.65	5.69	7.75	5.29	0.33
Cluster-3	26.71	14.20	3.70	8.93	6.35	0.40
Cluster-4	10.65	12.60	3.08	8.53	2.38	0.13
Cluster-5	22.70	26.09	8.48	8.52	8.72	0.70
Cluster-6	2.27	4.88	2.14	13.41	2.89	0.10
Cluster-7	1.78	12.19	3.84	12.51	3.13	0.17
Cluster-8	19.62	18.66	5.98	12.53	8.60	0.52
Cluster-9	78.14	24.81	3.13	8.09	3.45	0.20

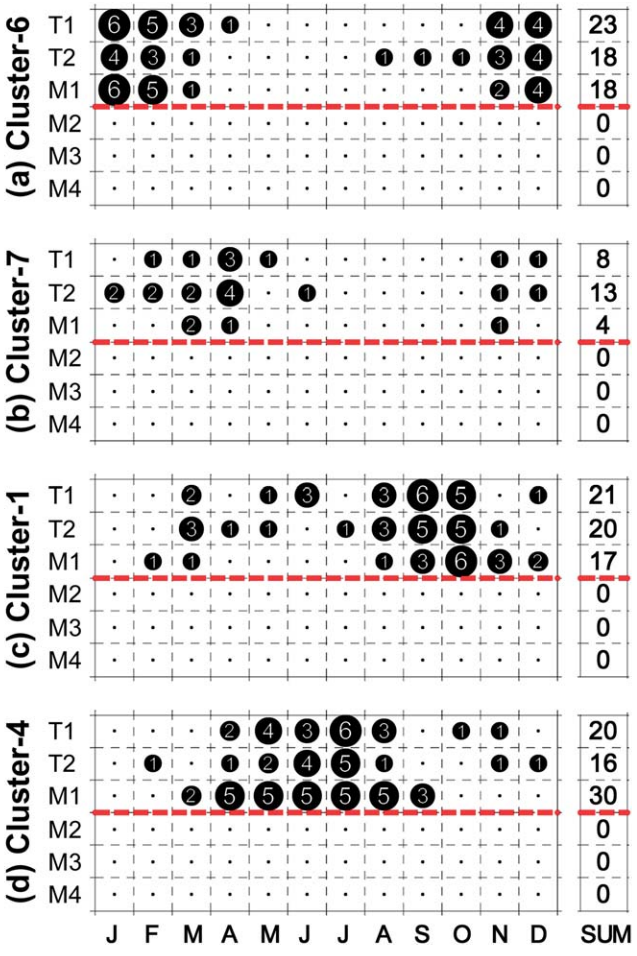

Fig. 8 Spatiotemporal meshes for (a) cluster-6, (b) cluster-7, (c) cluster-1 and (d) cluster-4 in the respective stations on a monthly basis with the thick dashed lines representing the conceptual location of the Gwangju metropolitan city and the sums for each station on the right.

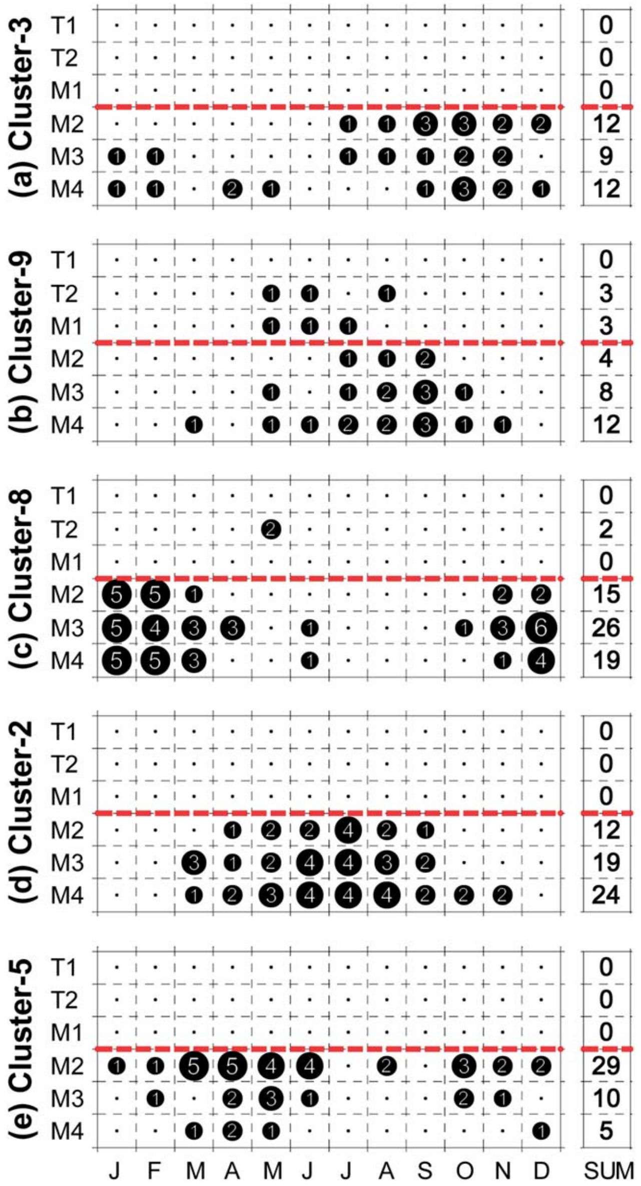

Fig. 9 Spatiotemporal meshes for (a) cluster-3, (b) cluster-9, (c) cluster-8, (d) cluster-2 and (e) cluster-5 in the respective stations on a monthly basis with the thick dashed lines representing the conceptual location of the Gwangju metropolitan city and the sums for each station on the right.

indicating the worst water quality condition. It should be noted that there is a drastic deterioration of water quality between M1 and M2 stations where the Gwangju metropolitan city is worsening the water quality in the river.

Concerning temporal variations, the environmental monitoring data associated with cluster-6, representing the best water quality condition, were mainly measured during winter as shown in Fig. 8(a). Cluster-7 and 1 (Fig. 8(b) and (c)) mainly contain data measured during autumn and spring, but the data for cluster-4 (Fig. 8(d)) with relatively low DO were measured during summer. The data classified into cluster-5, representing the worst water quality, were mostly measured during spring (Fig. 9(e)). The data in cluster-3 and 8 (Fig. 9(a) and (c)) were observed during autumn and winter, while cluster-9 and 2, with significantly high runoff and extremely low DO, had a high frequency mainly in summer as shown in Fig. 9(b) and (d). The temporal distribution of the data revealed that the clusters including the data measured during spring and summer generally showed worse water quality conditions with low DO values, due to a seasonal effect related to high temperature.

Characterizing the spatiotemporal variation of each cluster in detail, cluster-6 includes the runoff and water quality measured only in the T1, T2 and M1 stations during winter season and cluster-7 was related to the data measured in the same three upstream stations during spring and winter. The data measured in the same stations were classified into cluster-1 for autumn and cluster-4 for summer.

Cluster-5 contains the runoff and water quality measured only in M2, M3 and M4 stations mainly during spring. The data measured during autumn were associated with cluster-3 and the data for spring were classified into cluster-8 in the downstream three stations. Cluser-9 has the data measured during summer mainly in the downstream stations with six observations in T2 and M1 stations, while cluster-2 includes the summer data measured only in the downstream three stations.

The spatiotemporal mesh analysis using the results from pattern classification by SOM application clearly showed that the river basin may be divided into two areas with different water quality patterns which are influenced by the seasonal effect. The analysis confirmed that the SOM application classifying the parameters into the nine clusters was reasonable and feasible for the river basin. It also summarized the spatiotemporal distribution of the respective nine clusters with the readily understandable visualization. That is, relatively better water quality conditions were found in the three upstream stations whereas the three downstream stations showed worse water quality conditions. The water quality in spring and summer was generally worse than in autumn and winter in both areas. The spatiotemporal mesh analysis proposed in the present study was thus found useful for characterizing and understanding spatial and temporal variability and interdependence of runoff and water quality parameters measured in multiple stations.

Conclusions

In the present study, a Self-Organizing Map (SOM) combined with a hierarchical cluster analysis was applied for pattern classification of environmental monitoring data from the Yeongsan River basin in Korea, including water quality parameters and runoff measured in six stations. The SOM was systematically applied with a step-wise procedure including data transformation, determination of the SOM structure, initialization of reference vectors, training with relevant parameters, selection of an optimal number of clusters and a fine-tuning cluster analysis. The first, second and third quartiles of the reference vectors were plotted on radar charts to display fundamental characteristics of each cluster. In addition, the number of data occurrences in the respective stations on a monthly basis for each cluster was displayed in spatiotemporal meshes in order to characterize the spatiotemporal variability of the environmental monitoring data.

The spatiotemporal distribution of the environmental monitoring data was examined based on the characteristics of respective clusters. The spatial distribution revealed that the water quality condition was generally better upstream than downstream. The temporal distribution showed a clear seasonal effect. The best water quality conditions were associated with data measured in the upstream part of the basin during winter, while poor water quality conditions were found in the clusters with low DO measured in the downstream part during spring and summer.

To conclude, by the systematic application of a SOM, it was possible to classify the overall environmental aspect into a number of characteristic patterns with exclusively distinguishable environmental conditions. Therefore, it has proved to be practically applicable for assessment of the relative impact of the respective sub-basins on the overall environmental condition in the river basin represented by the runoff and water quality parameters. Specifically, the area downstream of the Gwangju metropolitan city associated with poor water quality conditions should be prioritized when designing and implementing environmental measures for comprehensive water quality improvement and watershed management. In addition, based on the applicability and feasibility of the SOM shown in this study, SOMs are expected to be utilized for integrated assessment of a river basin with simultaneous consideration of ecological, environmental and geographical factors.

Acknowledgements

The present study was carried out as a part of the research project “Solutions for the water related problems in Asian Metropolitan areas”, supported by the Tokyo Metropolitan Government, Japan.

References

1 A. Baker, *Hydrol. Processes*, 2003, **17**, 2499–2501.

2 J.-H. Kang, S. W. Lee, K. H. Cho, S. J. Ki, S. M. Cha and J. H. Kim, *Water Res.*, 2010, **44**, 4143–4257.

3 Ministry of Environment, *ECOREA, Environmental Review 2009*, Korea, 2009.

4 Y. J. Jung, M. K. Stenstrom, D. I. Jung, L. H. Kim and K. S. Min, *Desalination*, 2008, **226**, 97–105.

5 National Institute of Environmental Research, *Sustainable Promise, 2009 Annual Report*, Korea, 2009.

6 H. Chang, *Water Res.*, 2008, **42**, 3285–3304.

7 H. W. Lee, K. J. Bhang and S. S. Park, *Ecol. Inf.*, 2010, **5**, 281–292.

8 S. Shanmuganathan, P. Sallis and J. Buckeridge, *Environ. Modell. Software*, 2006, **21**, 1247–1256.

9 M.-Y. Song, H.-J. Hwang, I.-S. Kwak, C. W. Ji, Y.-N. Oh, B. J. Youn and T.-S. Chon, *Ecol. Modell.*, 2007, **203**, 18–25.

10 Y.-S. Park, M.-Y. Song, Y.-C. Park, K.-H. Oh, E. Cho and T.-S. Chon, *Ecol. Modell.*, 2007, **203**, 26–33.

11 L. Zhang, M. Scholz, A. Mustafa and R. Harrington, *Bioresour. Technol.*, 2009, **100**, 559–565.

12 D. Bedoya, V. Novotny and E. S. Manolakos, *Ecol. Modell.*, 2009, **220**, 2393–2406.

13 L. Faggiano, D. Zwart, E. García-Berthou, S. Lek and M. Gevrey, *Sci. Total Environ.*, 2010, **408**, 2319–2326.

14 K.-S. Jeong, D.-G. Hong, M.-S. Byeon, J.-C. Jeong, H.-G. Kim, D.-K. Kim and G.-J. Joo, *Ecol. Inf.*, 2010, **5**, 293–303.

15 L. E. Besaw, D. M. Rizzo, M. Kline, K. L. Underwood, J. J. Doris, L. A. Morrissey and K. Pelletier, *J. Hydrol.*, 2009, **373**, 34–43.

16 A. Hentati, A. Kawamura, H. Amaguchci and Y. Iseri, *Geomorphology*, 2010, **122**, 56–64.

17 K. L. Hsu, H. V. Gupta, S. Sorooshian and B. Imam, *Water Resour. Res.*, 2002, **38**, 1–38.

18 S. Srinivasulu and A. Jain, *Appl. Soft Comput.*, 2006, **6**, 295–306.

19 A. Jain and S. Srinivasulu, *J. Hydrol.*, 2006, **317**, 291–306.

20 H. Moradkhani, K. L. Hsu, H. V. Gupta and S. Sorooshian, *J. Hydrol.*, 2004, **295**, 246–262.

21 G.-F. Lin and L.-H. Chen, *J. Hydrol.*, 2006, **324**, 1–9.

22 A. M. Kalteh, P. Hjorth and R. Berndtsson, *Environ. Modell. Software*, 2008, **23**, 835–845.

23 R. Céréghino and Y.-S. Park, *Environ. Modell. Software*, 2009, **24**, 945–947.

24 H.-C. Lu, J.-C. Hsieh and T.-S. Chang, *Atmos. Res.*, 2006, **81**, 124–139.

25 K. Nishiyama, S. Endo, K. Jinno, C. B. Uvo, J. Olsson and R. Berndtsson, *Atmos. Res.*, 2007, **83**, 185–200.

26 H. L. Garcia and I. M. González, *Eng. Appl. Artif. Intell.*, 2004, **17**, 215–225.

27 Ö. Çinar, *Process Biochem.*, 2005, **40**, 2980–2984.

28 P. A. Aguilera, A. G. Frenich, J. A. Torres, H. Castro, J. L. M. Vidal and M. Canton, *Water Res.*, 2001, **35**, 4053–4062.

29 M. Alvarez-Guerra, C. González-Piñuela, A. Andrés, B. Galán and J. R. Viguri, *Environ. Int.*, 2008, **34**, 782–790.

30 S. M. Mustonen, S. Tissari, L. Huikko, M. Kolehmainen, M. J. Lehtola and A. Hirvonen, *Water Res.*, 2008, **42**, 2421–2430.

31 M. Tobiszewski, S. Tsakovski, V. Simeonov and J. Namiesnik, *Chemosphere*, 2010, **80**, 740–746.

32 S. Su, J. Zhi, L. Lou, F. Huang, X. Chen and J. Wu, *Phys. Chem. Earth*, 2010, **36**, 379–386.

33 A. Astel, S. Tsakovski, V. Simeonov, E. Reisenhofer, S. Piselli and P. Barbieri, *Anal. Bioanal. Chem.*, 2008, **390**, 1283–1292.

34 S. Tsakovski, A. Astel and V. Simeonov, *J. Chemom.*, 2010, **24**, 694–702.

35 S. J. Ki, Y. G. Lee, S.-W. Kim, Y.-J. Lee and J. H. Kim, *Water Sci. Technol.*, 2007, **55**, 367–374.

36 J. Vesanto, J. Himberg, E. Alhoniemi and J. Parhankangas, *SOM Toolbox for Matlab 5*, Helsinki University of Technology Report A57, 2000.