Available online at www.sciencedirect.com

ScienceDirect

Environmental Modelling & Software 23 (2008) 835–845

Environmental Modelling & Software

www.elsevier.com/locate/envsoft

Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application

A.M. Kalteh*, P. Hjorth, R. Berndtsson

Department of Water Resources Engineering, Lund University, P. O. Box 118, S-221 00 Lund, Sweden

Received 13 October 2006; received in revised form 3 October 2007; accepted 4 October 2007
Available online 19 November 2007

Abstract

The use of artificial neural networks (ANNs) in problems related to water resources has received steadily increasing interest over the last decade or so. The related method of the self-organizing map (SOM) is an unsupervised learning method to analyze, cluster, and model various types of large databases. There is, however, still a notable lack of comprehensive literature review for SOM along with training and data handling procedures, and potential applicability. Consequently, the present paper aims firstly to explain the algorithm and secondly, to review published applications with main emphasis on water resources problems in order to assess how well SOM can be used to solve a particular problem. It is concluded that SOM is a promising technique suitable to investigate, model, and control many types of water resources processes and systems. Unsupervised learning methods have not yet been tested fully in a comprehensive way within, for example water resources engineering. However, over the years, SOM has displayed a steady increase in the number of applications in water resources due to the robustness of the method.

Keywords: Artificial neural networks; Self-organizing map; Review; Water resources

1. Introduction

Modelling of hydrological processes that are embedded with high complexity, dynamism, and non-linearity in both spatial and temporal scales is of prime importance for hydrologists and water resources engineers. In many cases, however, the lack of physical understanding of the complex processes involved creates problems to find efficient models. Over the last decades artificial neural networks (ANNs) have been subject to an increasing interest in water resources problems. This has led to a tremendous surge in research activities (ASCE, 2000b; Maier and Dandy, 2000; Dawson and Wilby, 2001; Alp and Cigizoglu, 2007; Darsono and Labadie, 2007; Iliadis and Maris, 2007; Raduly et al., 2007). The increasing number of applications of ANNs in modelling of hydrological processes is related to their ability to relate input and output variables in complex systems without any requirement of a detailed understanding of the physics of the process involved (Dawson and Wilby, 2001). According to ASCE (2000a,b), an ANN is a massively parallel-distributed information processing system resembling biological neural networks of the human brain and capable of solving large-scale complex problems such as pattern recognition, non-linear modelling, classification, and control. The feed-forward multi-layer perceptron (MLP) is the most widely used ANN for prediction and forecasting of water resources variables (Maier and Dandy, 2000). Detailed reviews of ANNs along with assessments of their application in water resources and hydrology can be found in Maier and Dandy (2000), ASCE (2000a,b), and Dawson and Wilby (2001). The self-organizing map (SOM; also called Kohonen map or topology preserving feature map) is a kind of ANN method which is capable of clustering, classification, estimation, prediction, and data mining (Alhoniemi et al., 1999; Vesanto and Alhoniemi, 2000; Kohonen, 2001) in a wide-spread range of disciplines regarding signal recognition, organization of large collections

* Corresponding author. Present address: Department of Forestry, Faculty of Natural Resources, Guilan University, P.O. Box 1144, Sowmehe Sara, Guilan, Iran.

E-mail address: aman_mohammad.kalteh@tvrl.lth.se (A.M. Kalteh).

1364-8152/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2007.10.001

of data, process monitoring and analysis, and modelling as well as water resources problems. Typical for an SOM is that the desired solutions or targets are not given and the network intelligently learns to cluster the data by recognizing different patterns.

Despite the rather broad existing literature about ANN methods, in particular feed-forward MLPs (i.e., Maier and Dandy, 2000; ASCE, 2000a,b; Dawson and Wilby, 2001), there is a notable lack of comprehensive literature review on the efficiency of unsupervised learning techniques. Consequently, the main objective of this paper is to explain the SOM algorithm and to review the successes or failures of published applications with main emphasis on water resources and related disciplines. The paper is organized into two main parts. In the first part, the feed-forward MLP and SOM methods are explained along with a presentation of their structural differences. In the second part, published applications of the SOM method in water resources problems and related disciplines are reviewed and evaluated. We close the paper by giving some future avenues for the application of SOMs in water resources.

2. Artificial neural networks (ANNs)

2.1. Feed-forward multi-layer perceptron (MLP)

As stated above, the most commonly used ANN in water resources and hydrology is the feed-forward MLP as shown in Fig. 1. In this figure, each neuron is represented by a circle and each connection weight by a line, and the structure of an individual neuron is shown. Each individual neuron computes an output, based on the weighted sum of all its inputs, according to a non-linear function called the activation function such as the hyperbolic tangent(see Fig. 1 and Eq. (1)):

$$f(x) = \frac{2}{1+e^{-2(x)}} - 1 \tag{1}$$

or the sigmoid (see Eq. (2)) activation function:

$$f(x) = \frac{1}{1+e^{-x}} \tag{2}$$

In Eqs. (1) and (2), x is the weighted sum of inputs to the neuron and $f(x)$ is the neuron's output.

Dawson and Wilby (2001) claim that sigmoid and hyperbolic activation functions are the most common ones and that they are used in the majority of applications. The feed-forward MLP shown in Fig. 1 consists of three layers: an input layer consisting of input neurons where the number of neurons is equal to the number of explanatory variables, a hidden layer where the number of neurons is usually chosen via a trial-and-error procedure, and an output layer where the number of neurons is equal to the number of output variables. There is also a bias neuron in each of the hidden and output layers (as shown in Fig. 1). Feed-forward MLPs must be trained by means of training algorithms that create learning processes serving to find an optimal set of weights for the connections and bias values for the neurons. Back-propagation is the most popular algorithm for training feed-forward MLP. In the feed-forward MLP, patterns from the inputs presented to the neurons in the input layer are propagated through the network from the input layer to the output layer, i.e. in a forward direction and the outputs from the network are compared with the target values in order to compute the error. Thereafter the calculated error is back propagated through the network and the connection weights and bias values are updated (ASCE, 2000a). This process is repeated until convergence. When the training has been finished, the network is able to produce outputs given unseen inputs. For detailed reviews of ANN training algorithms along with their application in water resources and hydrology we refer to Maier and Dandy (2000), ASCE (2000a,b), and Dawson and Wilby (2001).

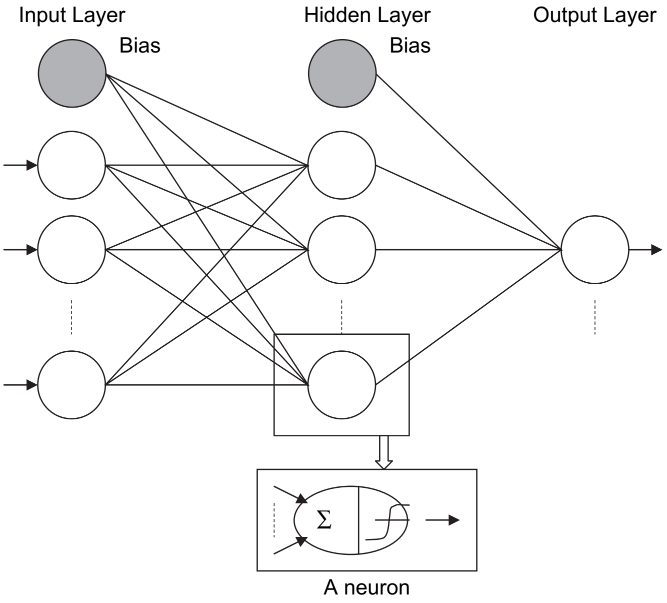

Fig. 1. Structure of a feed-forward multi-layer perceptron (MLP) (modified from Kalteh and Berndtsson, 2007).

2.2. Self-organizing map (SOM)

The self-organizing map (SOM) is a learning algorithm that was originally proposed by Kohonen (1982a,b). The SOM is a fascinating neural network method that has found increasing interest in water resources applications such as, e.g., classification of satellite imagery data and rainfall estimation (Murao et al., 1993) and rainfall–runoff modelling and analysis (Hsu et al., 2002). Typically, SOM networks learn to cluster groups of similar input patterns from a high dimensional input space in a non-linear fashion onto a low dimensional (most commonly two-dimensional) discrete lattice of neurons in an output layer (Kohonen, 2001). This is done in such a way that neurons physically located close to each other in the output layer of the SOM have similar input patterns (combining clustering and ordering processes in SOM). Higher dimensional output layers are also possible, but they will not be so convenient for visualization purposes and consequently they are not so common (Vesanto, 1999). Discrete lattices can be either hexagonal or rectangular but hexagonal are preferred because they are effective for visualization and/or more convenient to the eye (Vesanto, 1999; Kohonen, 2001). The main advantages of the SOM algorithm are that it is non-linear and has an ability to preserve the topological structure of the data (Corne et al., 1999; ASCE, 2000a). In general, the SOM algorithm clusters the samples or patterns into predefined (i.e. the number of neurons is selected by the modeller) classes and also orders the classes into meaningful maps (topology preservation or ordering property). The typical structure of an SOM consists of two layers: an input layer and a Kohonen or output layer (Fig. 2). The input layer contains one neuron for each variable (e.g., precipitation, temperature, etc.) in the data set. The Kohonen layer neurons are connected to every neuron in the input layer through adjustable weights or network parameters. The weight vectors in the Kohonen layer give a representation of the distribution of the input vectors in an ordered fashion.

The successive procedures required to apply SOM can be divided into three categories, namely:

(i) *Data gathering and normalization*: the most important part in normalizing is to prevent variables from having higher impact as compared to other variables. Consequently, normalization, by transforming all the variables to the range of e.g., 0–1, ensures that all variables have equal importance in the formation of the SOM.

(ii) *Training*: after data preparation and normalization, an input vector from the data matrix is introduced to the iterative training procedure to form the SOM. It is recommended that the number of iterations should be at least 500 times the number of neurons in the output layer (Haykin, 1999; Kohonen, 2001). At the outset of training, weight vectors must be initialized by using either a random or a linear initialization method. The weight vectors are also called reference or codebook vectors. Random initialization of weight vectors is the most commonly used method in hydrological applications. Whatever type of initialization method that is applied, the SOM utilizes a type of learning that is called competitive, unsupervised, or self-organizing procedure to match each input vector with a neuron in the SOM. This is done by comparing the presented input pattern of a data matrix with each of the SOM neuron weight vectors. The neuron with the closest match to the presented input pattern is called winner neuron or best matching unit (BMU). The most common criterion that is applied to find the winner neuron is Euclidean distance. Then, the weight vector of the BMU and the topologically neighboring neurons are updated in such a way as to reproduce the input pattern. The most commonly used neighborhood function is the Gaussian:

$$N_{j^*j}(t) = \mathrm{e}^{-\frac{\left\|r_{j^*} - r_j\right\|^2}{2\sigma^2(t)}} \tag{3}$$

where $N_{j^*j}(t)$ is the neighborhood function of the best matching neuron j^* at iteration t; $\delta(t)$ is the neighborhood radius at iteration t; and $\|r_{j^*} - r_j\|$ is the distance between neurons j^* and j on the map grid.

This process is repeated until convergence.

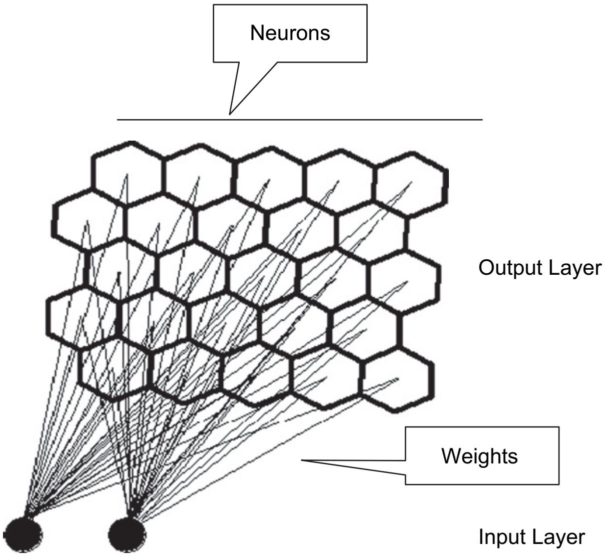

Fig. 2. Structure of a 5 × 5 two-dimensional self-organizing map (SOM) (modified from Kalteh and Berndtsson, 2007).

(iii) *Extracting information from the trained SOM*: once the training of the SOM has been accomplished the resulting map can be post-processed based on visualization, clustering or local modelling purposes. This issue has been comprehensively investigated by Vesanto (2002). From the point of view of a water resources engineer or a hydrologist, the trained SOM map is a valuable tool for visualization of the relatively large amount of data along with getting insight into the system under investigation such as precipitation processes (e.g., Kalteh and Berndtsson, 2007) or rainfall–runoff processes (e.g., Hsu et al., 2002). As mentioned above, the SOM is an unsupervised clustering algorithm and in most applications reviewed in this paper it is used for this purpose. However, we will concentrate on innovative and creative applications of SOM in water resources and hydrology. An example of SOM clustering is shown in Fig. 3. As seen from the figure, the first-level SOM clusters input data into a given number of clusters and the second-level clusters the output neurons into as many different regions as required.

The issues regarding data gathering, and normalization, training and post-processing are well-documented in Vesanto (1999), Vesanto et al. (2000), Kohonen (2001), and Vesanto (2002) hence no details are given here.

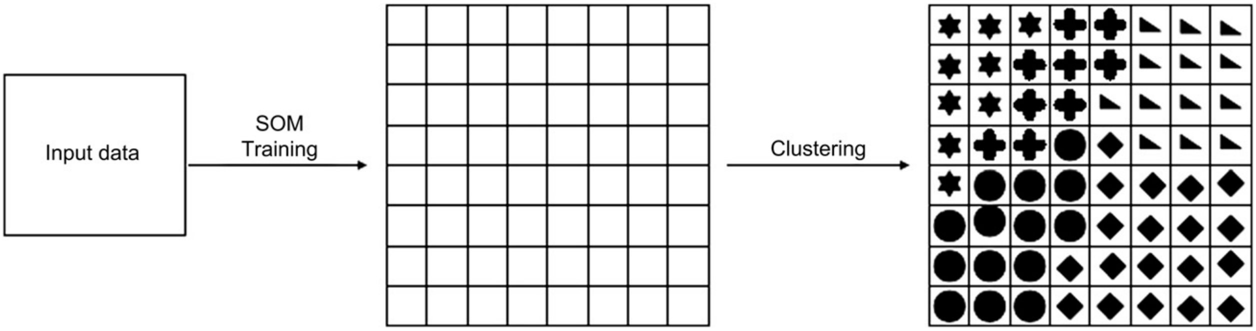

Fig. 3. The diagram of a two-level clustering of the SOM. Different symbols represent different clusters (Wu and Chow, 2004: with kind permission of Elsevier).

3. Applications in water resources and hydrology

In the former section, we briefly discussed the SOM structure, and its basic concepts along with procedures required to apply the SOM algorithm to a data set. In this section, we review some successful SOM applications with emphasis on innovative and creative solutions for analysis, estimation and prediction of various hydrological processes such as precipitation, river flow, rainfall–runoff, surface water quality, and other related disciplines such as climate and environment.

3.1. SOM in river flow and rainfall–runoff

Modelling the rainfall–runoff relationship in a watershed is of prime importance to water resources engineers and hydrologists in design of hydraulic structures, flood control and management. This relationship is believed to be highly complex, dynamic and non-linear in both spatial and temporal scales. Due to the robustness of ANNs many researchers have used ANNs for modelling rainfall–runoff relationships (e.g., Maier and Dandy, 2000; ASCE, 2000b; Dawson and Wilby, 2001). However, existing review papers mainly consider ANNs other than SOM which is the main objective of this review. Hence, this study aims at filling a gap by addressing the potential of SOM in modelling watershed runoff.

In a preliminary study, Hall and Minns (1999) used an SOM algorithm for regionalization of gauging sites in southwestern England and Wales based on five catchment characteristics per gauging site, i.e., catchment area, main stream length, main stream slope, mean annual rainfall, and winter rain acceptance potential or soil index. The authors decided to use 10 neurons in the output layer of an SOM since they had an expectation of at least two, or possibly three, classes. As discussed earlier, once an SOM is trained it will map gauging sites with similar characteristics to the same neurons in the output layer. Thus, Hall and Minns (1999) grouped the output neurons into three distinct groups in order to obtain three homogeneous regions. The study showed the potential of SOM to group data into homogeneous areas, which is useful where there are needs to transfer information from gauged to ungauged sites, a common problem for geographically remote areas and developing countries. More recently, Lin and Chen (2006) used an SOM to group 154 rain gauges in Taiwan for regional frequency analysis. The input variables were: gauge latitude (m), gauge longitude (m), elevation (m), mean annual rainfall (mm), standard deviation of annual rainfall (mm), and mean monthly rainfall (mm) for each month (totally 17 variables). The authors decided to use a map size of 12 × 12 in order to ensure that the maximum number of clusters (classes) would be obtained from the training data. This was a rather arbitrary decision since there is no theoretical principle to determine the optimum number of neurons in the output layer of an SOM. They obtained eight homogeneous regions by dividing the output layer neurons of the SOM into eight clusters in such a way that the borders between them were made up of neurons without any patterns projected to them. They compared these clustering results with two other conventional clustering methods, i.e., the *k*-means and Ward's methods and found that the SOM is able to identify homogeneous regions more precisely than the other two methods.

A study by Furundzic (1998) concerning the use of an SOM to decompose the rainfall–runoff process input–output space into three classes and thereafter, employing a separate feed-forward MLP model for each class, opened up several possibilities for multinetwork modelling approaches with decomposition of the modelling domain. Abrahart and See (2000) compared the forecasting power of ANN and autoregressive moving average (ARMA) models in two different catchments in the UK. They found that the models produce similar results. They also used an SOM to cluster the modelling domain into distinct individual event types. The input variables to the SOM were 6 h of previous flow data. After examination of map sizes – 2 × 2, 4 × 4, 6 × 6 and 8 × 8 – using various data sets, the authors decided to use a map size of 8 × 8 or 64 clusters for the output layer of the SOM. In order to examine the potential merits of this approach to clustering of the modelling domain, a separate ANN model was developed for each of the two most prevalent rising event clusters for each individual station (Kilgram, Skelton, and Cefn Brwyn) in the catchments of River Ouse and Upper River Wye, respectively. This was found to produce improved modelling performance.

Hsu et al. (2002) developed a Self-Organizing Linear Output mapping network (SOLO) for hydrological modelling and analysis. The SOLO consists of an input layer, an input classification layer that is obtained by using an SOM, and a mapping layer that maps the inputs to the outputs using piecewise linear regressions. They applied the SOLO ANN architecture, to forecast one-day ahead stream flow for the Leaf River basin (1949 km^2) near Collins, Mississippi. Rainfall and runoff data for the three previous days were used as input variables to the network.

The SOLO model uses a data classification scheme and piecewise linear regressions such that once data classification is achieved by SOM a linear regression function is fitted to the data included in each neuron. They arbitrarily selected map sizes of 2×2 and 15×15 for illustration of the underlying structure of the input–output process in the SOM layer and for exploration of detailed results, respectively. However, in order to determine the optimal map size, they conducted a series of experiments using progressively larger map sizes. The root mean square error, correlation and bias statistics were evaluated by the authors as a function of map size and they only found marginal improvements for map sizes exceeding 5×5. The authors compared the performance of the SOLO model with an autoregressive model with exogenous inputs (ARX), a multi-layer feed-forward network (MFN), a recurrent neural network (RNN) and the Sacramento soil moisture accounting (SAC-SMA) model. Fig. 4 shows the performance of these models for the highest validation flow year (1980). As seen the SOLO and MFN models track all portions of the hydrograph more closely. Based on their study, Hsu et al. (2002) concluded that the SOLO model not only provides better predictions but its classification layer also provides insight into the system under investigation. To be fair to the SAC-SMA model, it should be noted that it is based on fewer input variables and also provides a conceptual model of the physical processes involved.

In a related study dealing with multiple models and modular learning for modelling hydrological processes through decomposition of the modelling domain, Parasuraman et al. (2006) developed spiking modular neural networks (SMNNs) for modelling hydrological processes, based on the concepts of both self-organizing networks and modular networks. The SMNN consists of an input layer, a spiking layer that clusters the input, and an associator neural networks layer which is composed of MLP models to associate the clustered input patterns to outputs. Classification of the input space in the spiking layer is achieved by means of (1) competitive learning and (2) SOMs, such that the former learns the distribution and the latter learns both the distribution and the topology of the input space. Once the classification of the input space is achieved, mapping of inputs to corresponding outputs is achieved by feed-forward MLP models in the associator neural networks layer. The configuration of the SMNNs is shown in Fig. 5. The authors evaluated the performance of the proposed model in modelling stream flow and eddy covariance-measured evapotranspiration, respectively. The stream flow model used the monthly stream flow values of the English River, Ontario, Canada, between Umfreville and Sioux Lookout such that the monthly stream flow at Umfreville was used as input variable in order to estimate stream flow values at Sioux Lookout. The evapotranspiration model used air temperature, ground temperature, net radiation, relative humidity, and wind speed as input variables in order to model hourly latent heat flux. The authors decided to use two and eight clusters, determined by a trial-and-error procedure such that by starting with two neurons, different numbers of neurons were evaluated with the objective of minimization of a cost function. It was found that the SMNNs performed better than a single feed-forward MLP. The authors claimed that the SMNNs were able to decompose the complex processes effectively into simpler ones that can be learned with relative ease.

Jain and Srinivasulu (2006) presented a procedure for decomposing a flow hydrograph into different segments based on physical concepts in a catchment and thereafter modelling the different segments using feed-forward MLPs and conceptual techniques. They tested the proposed procedure for rainfall–runoff data in the Kentucky River catchment (17,820 km^2). Moreover, they developed one-dimensional SOM models for decomposing the effective rainfall–runoff data into different segments (3 and 4, respectively) to test the proposed procedure. They concluded that dividing the rainfall–runoff data into different segments based on the physical concepts was better than relying on the SOMs for classification. Considering the above studies one can argue that decomposition of the modelling domain into simpler ones and devoting a separate model to each sub-domain improves the model performance. These studies indicate that the SOM model is valuable for decomposition because it can be used for decomposition without prior knowledge about the hydrological processes involved. While traditional ANN models are considered as black-box models these studies may indicate that SOMs are not purely black-box models and that it is possible to get some insight into the processes being investigated. Still, however, there seems to be a great potential to make a contribution regarding the issue of advancing the modular learning of the processes.

Moradkhani et al. (2004) tried to enhance the modelling accuracy in stream flow forecasting by means of radial basis function (RBF) networks combined with SOMs. The structure of the RBF networks is identical to the feed-forward MLP (Fig. 1) in which Gaussian basis functions (Eq. (4)) are used in the hidden layer:

$$f(x) = \mathrm{e}^{-\frac{x^2}{2\sigma^2}} \qquad (4)$$

where x is the weighted sum of inputs to the neuron, σ is the radius of the basis function and $f(x)$ is the neuron's output.

They developed a so-called Self-Organizing Radial Basis (SORB) function in order to forecast one-step ahead daily stream flow in the semi-arid Salt River basin (10,000 km^2), a sub-regional watershed of the lower Colorado River basin in the United States. The SORB uses the Gaussian Radial Basis Function architecture in combination with an SOM so that the SOM was used to cluster the input data for extracting the Gaussian function parameters (center and spread parameters). The network parameters obtained by the SOM can be considered as center parameters. The initial spread parameters can be estimated by calculating the standard deviation of the

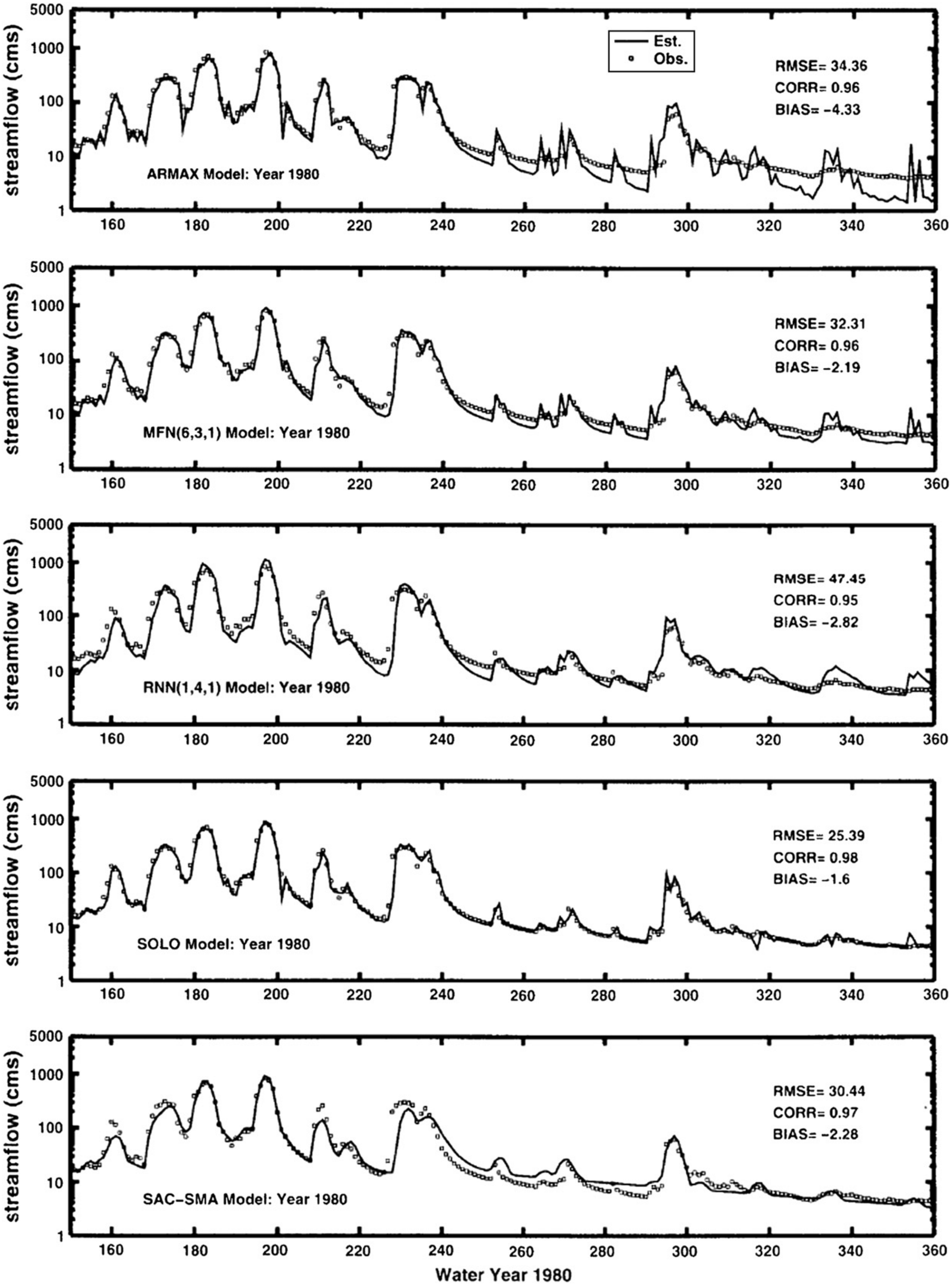

Fig. 4. The performance of testing models for the wettest year (1980) of the evaluation period (Hsu et al., 2002).

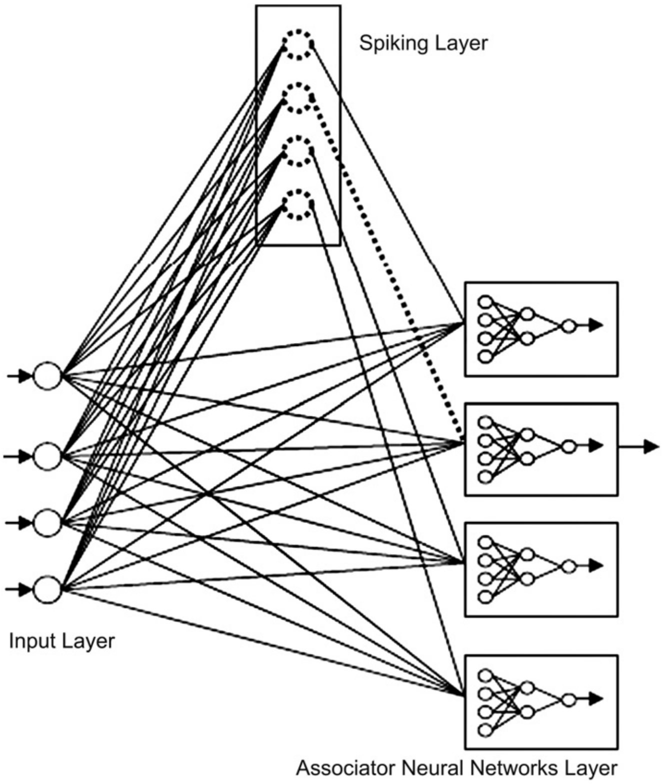

Fig. 5. Configuration of a spiking modular neural network (SMNN). The spiking layer is the equivalent of a classification layer (Parasuraman et al., 2006).

data points in each cluster and are to be fine-tuned in the training phase. Three previous days of stream flow, two previous days of precipitation, average precipitation in the period of 5–14 and 30–39 days in the past, and average temperature in the period of 1–5 days in the past were used as input variables. The authors compared the performance of SORB with feed-forward MLP, SOLO (Hsu et al., 2002), and linear regression (LINREG) models and found a relative superiority of SORB in terms of forecasting accuracy. Although the feed-forward MLP type of ANNs is the most commonly used one in water resources applications, there are many advantages in the RBF networks which make them attractive, such as faster convergence, higher reliability, and a good generalization ability with a minimum number of neurons. However, the positioning of the radial basis centers is a crucial obstacle in designing of the RBF networks, which can be solved by combining with SOM as in the above study.

3.2. SOM in precipitation

As seen from the discussion above, precipitation is a most important component of the rainfall–runoff process. Thus, it is required as input to these types of models. Precipitation itself, however, is difficult to simulate due to its large variability both in space and time. Many ANN applications for estimation of precipitation were comprehensively reported in ASCE (2000b). However, in this sub-section we review some published SOM applications, in order to clarify the potential benefits of SOMs in precipitation simulation studies. In a preliminary study, Murao et al. (1993) proposed and used a hybrid ANN consisting of an SOM for classification of textural feature vectors from satellite imagery data and multiple feed-forward MLP models for rainfall estimation with the same architecture as in Fig. 5. The results were encouraging. Hong et al. (2005) proposed and used a self-organizing nonlinear output (SONO) ANN architecture for estimation of rainfall based on cloud patch. The SONO is an extended SOLO model (described before) so that both models make use of an SOM in their classification layer. However, the SONO model maps input to the corresponding output by means of non-linear regression once classification of the input space is obtained by the SOM. To summarize, the SONO consists of a classification layer based on an SOM to classify the IR imagery into given cloud patches and a non-linear regression mapping layer as an approximation of different non-linear cloud-precipitation relationship mapping. The above studies used SOM for classification of input patterns and thereafter developed a separate model for each SOM class. However, extracting the information from an SOM can be done in a number of different ways to fulfill the need of water resources engineers and hydrologists. For example, Kalteh and Berndtsson (2007) used an SOM both for regionalization and for interpolation of monthly precipitation in northern Iran, a region with large complexity of precipitation mechanisms. They used an unsupervised SOM as a classifier for regionalization and thereafter a supervised SOM for interpolation. In the case of supervised SOM, some minor modifications were required in the training process so that finding the BMU was based only on the input portion of the data matrix presented to the SOM but updating applied to all input–output variables. The authors also used a component plane visualization technique in order to depict similar precipitation stations. This visualization technique is often used to visualize variables in such a way that precipitation stations that exhibit similar patterns can be recognized (Vesanto, 2002).

3.3. SOM in surface water quality

The non-linearity and complexity of variables involved in water quality have led many researchers to use ANN models to simulate these variables due to the ability of such models to handle complex, non-linear relationships. Many such ANN applications in this context have been documented in ASCE (2000b). In this sub-section, we review some studies in water quality context that use SOM to solve particular problems.

The issue of dividing data into training, testing, and validation subsets in ANN models development was addressed by Bowden et al. (2002). This issue is one of the most important in the development of ANN models. It needs to be carried out in an efficient way due to the fact that the division of data into subsets can have significant influence on the performance of the obtained ANN model. As ANN models learn from examples in training data, the model will have poor generalization ability if the training data are not representative of the modelling domain. Then, the model will be facing examples that are beyond the training set. And also, the inclusion of too many repetitive examples in a training set will only slow down the training procedure. Consequently, there is a need for a principled procedure to divide the data into subsets rather than to arbitrarily divide the data into subsets without consideration of the statistical properties, which is the most common procedure in current applications. Bowden et al. (2002) presented two methods based on a genetic algorithm and an SOM which were able to divide the available data into subsets with common statistical properties. These were used to develop back-propagation ANN models for salinity forecasting in the River Murray at Murray Bridge, South Australia, 14-days in advance. They compared the performance of the obtained models with a model that was developed by means of arbitrary division of data. They found that the models developed using the presented techniques outperformed the conventional one. It was found that the SOM also could be used as a tool to find out why the ANN models can give poor results in some parts of the time series. To this end, they utilized an SOM for clustering the data. An SOM with 10 × 10 neurons in the Kohonen layer was used so that the input patterns were clustered into 100 clusters, where 49 consisted of three or more patterns. From each of these clusters three data patterns were sampled, one each for training, testing, and validation. For the 51 clusters containing less than three patterns, they used the sample record in the cluster with only one record as training and in the case of clusters with two records, one was used for training and the other for testing.

Another issue that also needs great attention in the development of ANN models is the determination of significant input variables. This is due to the fact that the presentation of all potential input variables to the ANN and relying on the network to identify the most critical ones may create problems. As Bowden et al. (2005a,b) pointed out, there are several disadvantages with this approach, including the increase of computational complexity and memory requirements, difficulty in learning, misconvergence and poor model performance, increase of the complexity of the model and consequently, a difficulty in understanding the model as well as increasing noise due to inclusion of spurious input variables. Consequently, there is a need to introduce an efficient input determination method to select a more parsimonious model. This issue was addressed by Bowden et al. (2005a,b) who presented two methodologies, one based on the partial mutual information (PMI) algorithm and the other based on an SOM combined with a hybrid genetic algorithm and a general regression neural network (SOM-GAGRNN). The first uses a partial measure of the mutual information criterion in order to determine the inputs that have highly significant relationship with the output variable of the system being modelled and the second uses an SOM to cluster the input variables into groups of similar inputs. Then they selected one input variable from each cluster so that the input with the smallest Euclidean distance to the cluster's weights was selected from each cluster. Thereby the dimensionality of the input space was reduced and the variables obtained were introduced to the GAGRNN to determine what inputs that have significant relationship with the output variable of the system being modelled. The authors tested the proposed methods on several synthetic data sets and concluded that in terms of predictive performance, both methods were good while from the point of view of getting valuable information about the system under investigation the first was recommended. To verify the above approaches on real data, they were applied for input determination for an ANN model in forecasting salinity 14-days in advance in the River Murray at Murray Bridge, South Australia. The input variables for the ANN included daily salinity, flow, and river level data at 16 locations along the river. With 60 lags, there were, thus, 960 input variables. The proposed techniques were compared with two methods used in previous studies by Maier and Dandy (1996, 1997) for forecasting salinity. The authors found that the new techniques led to more parsimonious (in terms of number of inputs) ANN models and claimed that the models developed by means of the new techniques had higher generalization ability. Table 1 summarizes this comparison.

3.4. SOM in climate, environment, ecology, etc.

The SOM methodology has also been successfully applied in many climatic, environmental, and ecological applications. Chon et al. (1996) utilized an SOM for both clustering and patternizing community data in ecology. The input data patterns to SOM were the benthic macroinvertebrate communities collected at study sites in the Suyong River in Korea. The authors decided to use a 9 × 9 neurons map, without clarifying how and why, to cluster and patternize benthic macroinvertebrate communities. They found generally similar clustering results between SOM and conventional clustering analysis (the clustering based on the method of average linkage between groups). However, the authors claim that once the training of the SOM was accomplished, the most similar input patterns will be mapped into a specific neuron in such a way that each neuron in the map will respond to specific pattern(s) via searching for BMU. Consequently, when unseen input is presented to the map, it could be mapped into a particular neuron. The authors call this property patternizing, in differentiation from clustering. Thus, they claim this property is a basic difference between SOM and conventional clustering analysis

Table 1
Comparison of the best ANN models developed using the inputs obtained by Methods 1 (PMI) and 2 (SOM-GAGRNN) with the best ANN models developed using the inputs obtained in previous studies by Maier and Dandy (1996, 1997) (modified from Bowden et al., 2005b)

Input data set	Architecture	RMSE (EC units)		
		Training	Testing	Validation
Method 1	13–32–1	29.3	30.8	34.0
Method 2	21–33–1	30.5	38.0	36.2
Maier and Dandy (1996) (a priori + sensitivity)	39–30–1	43.0	40.3	43.0
Maier and Dandy (1997) (method of Haugh and Box)	47–35–1	43.9	38.2	46.2

even though they generally produce similar clustering results. Kothari and Islam (1999) used an SOM with 7 × 7 neurons in the Kohonen layer to characterize the spatial structure of remotely sensed soil moisture data from the little Washita Watershed located in southwest Oklahoma. They demonstrated that the results obtained were not very sensitive to the number of neurons selected. Corne et al. (1999) used an SOM in a supervised manner to model sub-glacial water pressure processes at Trapridge Glacier, Yukon Territory, Canada. After developing and evaluating a series of SOMs, the authors decided to use 8 × 8 neurons to perform this modelling because it performed reasonably well on the validation data. Cereghino et al. (2001) applied an SOM on ecological data that included the presence (1) or absence (0) (nominal data) of each species (283 species) at each site (252 sites) in the Adour–Garonne drainage basin (South-Western France) to classify the sampling sites as well as for visualization of the spatial distribution of each of the 283 considered species. The authors decided to use 10 × 15 neurons in the Kohonen layer and only claimed that the map size selection had been difficult, without providing further details. In a related study, Giraudel and Lek (2001) utilized an SOM for clustering sample units (in species abundance database which consists of the species and the sample units) and visualization of species abundance by utilizing the component planes visualization technique. They decided to use 4 × 4 neurons in the Kohonen layer and claimed that the selected map size is larger than the amount of sample units (there were 10 sample units). The authors compared the SOM with conventional techniques such as polar ordination, principal components analysis, correspondence analysis, and non-metric multi-dimensional scaling. They concluded that the SOM is useful in ecology and that it can serve as complementary technique for other conventional techniques. The utilization of a two-level SOM for clustering was addressed by Tran et al. (2003). They applied the SOM in environmental assessment to classify 123 watersheds in the Mid-Atlantic region into different clusters, in terms of environmental indicators (26 indicators), in combination with principal component analysis (PCA) which was applied to reduce the dimensionality of multivariate (indicators) data. The authors applied a two-level SOM for clustering the watersheds in such a way that the first-level SOM was applied directly to the data to create a 10 × 6 Kohonen map and then the second-level SOM was applied to the 60 prototype vectors of the first-level SOM to produce a 5 × 3 Kohonen map as shown in Fig. 6. The authors claimed that they have considered several factors in the selection of the number of neurons in the Kohonen layers (first- and-second-level SOMs) including the size of the data, possible clusters identified from the first-level SOM as well as consideration of having a reasonable number of clusters for later analysis. The issue of hybridization of SOM and RBF networks in ecological and environmental applications was addressed by Obach et al. (2001) who applied an SOM for clustering species abundance patterns. The authors also combined an SOM with a RBF network to predict aquatic insect abundance in such a way that the SOM' connection weights were used as centers of the RBF network. The authors decided to use 12 × 10 neurons for clustering and 6 × 1 neurons for prediction in the Kohonen layers in order to have a sufficient amount of data mapped on every single neuron. The authors claimed that the clustering results of SOM are comparable to statistical cluster analysis. Schütze et al. (2005) presented an ANN architecture based on SOM entitled self-organizing maps with multiple input–output (SOM-MIO), which after unique training, was able not only to perform a Richards equation simulation but also to produce its inverse solution with accuracy. The SOM-MIO combines the superior clustering capability of the SOM with a linear interpolation scheme (Delaunay interpolation scheme) to extract continuous output. The authors compared the SOM-MIO results with those obtained from a numerical model and found that the new architecture showed superior accuracy in both simulation and inverse solution tasks along with its computational efficiency. More recently, Shanmuganathan et al. (2006) discussed the immediate needs in data analysis, tools and modelling techniques to improve our understanding and prediction abilities of ecosystem response to the human influence for achieving sustainable environment development and management. In order to meet that, they used SOM methodologies for data analysis at regional scale (river water quality monitoring data from the Waikato River) and global scale (environmental and economic system data from the World Bank's

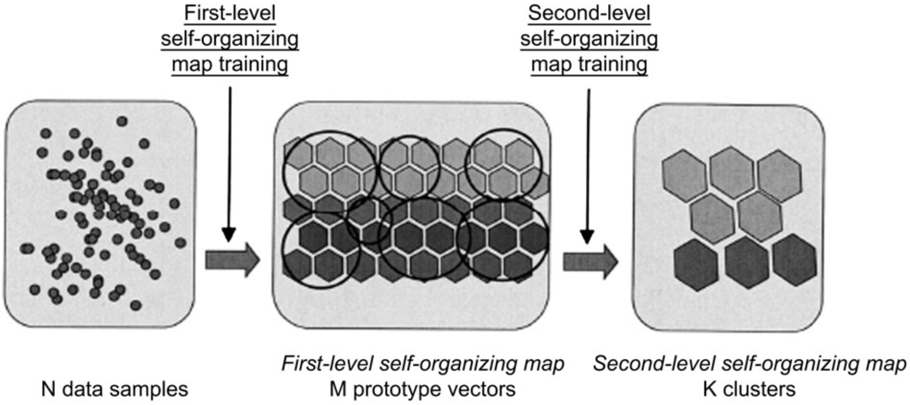
First-level self-organizing map training
Second-level self-organizing map training
N data samples
First-level self-organizing map
M prototype vectors
Second-level self-organizing map
K clusters

Fig. 6. The diagram representation of a two-level SOM clustering process (Tran et al., 2003: with kind permission of Springer Science and Business Media).

statistical tables). In the case of water quality data analysis, 19 parameters related to Waikato River water quality, sampled from different monitoring sites (1980–2002) were used as inputs to the SOM. The authors claim that SOMs are able not only to summarize a large amount of data but also to provide valuable information concerning the process under investigation via the visualization of changes in water quality parameters at different monitoring sites and relating it to human use. The latter was achieved by depicting the mapping of water quality parameters sampled at different sites on 4 June 2002, on the SOM in order to observe water quality changes as human use increases.

4. Conclusion and discussion

Over the last decades, SOMs have increasingly been used for analysis, estimation and prediction of various hydrological processes such as river flow, rainfall–runoff, precipitation, surface water quality, and related issues such as climate and environment. These studies indicate that in many cases, SOM can outperform other methods to solve various problems in water resources and hydrology. However, like feed-forward MLP applications, SOM applications are generally dependent on ad-hoc approaches characterized by guesswork and/or trial-and-error approaches. Likewise, there are no proven techniques to assess the reliability or validity of SOM or feed-forward MLP models.

Thus, there are areas that need further consideration. There is a need to further investigate the ability of SOMs to describe and to further analyze non-linear processes in water resources and to test the robustness of these methods in relation to traditional linear and/quasi-linear approaches. Examples of this are rainfall–runoff relationships and other types of hydrological processes often incorporating strong non-linearity. Hsu et al. (2002) provided some understanding of the input–output relationships by clustering of the input variables using SOM. However, it would be interesting to conduct this analysis in a more stringent way.

Another interesting ability of SOMs is that they can be used to automatically group and/or typify data according to different properties. Gaps in water resources data often create problems due to non-homogeneities and spatial heterogeneity. SOM may be used to find robust data-filling techniques that can help modellers and engineers in process control.

Even though SOMs have been used for modularization in modelling hydrological processes, there are still open issues to consider in this context concerning modularization in more principled ways. The multi-level SOM is a promising alternative approach to decomposition of the hydrological processes into simpler sub-domains.

Hydrologists are latecomers in the ANN field and have not caught up with the latest improvements in the field. Thus, there is much to be gained if hydrologists were able to apply recent advances concerning more principled model development and data preprocessing.

Still, successful applications of ANN methodologies in hydrology depend very much on the experience of the modeller. Therefore, it would be premature to present a simple collection of recipes that could be used by the practitioner.

References

Abrahart, R.J., See, L., 2000. Comparing neural network and autoregressive moving average techniques for the provision of continuous river flow forecasts in two contrasting catchments. Hydrological Processes 14, 2157–2172.

Alhoniemi, E., Hollmen, J., Simula, O., Vesanto, J., 1999. Process monitoring and modeling using the self-organizing map. Integrated Computer-Aided Engineering 6 (1), 3–14.

Alp, M., Cigizoglu, H.K., 2007. Suspended sediment load simulation by two artificial neural network methods using hydrometeorological data. Environmental Modelling and Software 22, 2–13.

ASCE Task Committee on application of Artificial Neural Networks in Hydrology, 2000a. Artificial neural networks in hydrology. I: preliminary concepts. Journal of Hydrologic Engineering 5 (2), 115–123.

ASCE Task Committee on application of Artificial Neural Networks in Hydrology, 2000b. Artificial neural networks in hydrology. II: hydrologic applications. Journal of Hydrologic Engineering 5 (2), 124–137.

Bowden, G.J., Dandy, G.C., Maier, H.R., 2005a. Input determination for neural network models in water resources applications. Part 1 – background and methodology. Journal of Hydrology 301, 75–92.

Bowden, G.J., Maier, H.R., Dandy, G.C., 2002. Optimal division of data for neural network models in water resources applications. Water Resources Research 38 (2), 1010, doi:10.1029/2001WR000266.

Bowden, G.J., Maier, H.R., Dandy, G.C., 2005b. Input determination for neural network models in water resources applications. Part 2. Case study: forecasting salinity in a river. Journal of Hydrology 301, 93–107.

Cereghino, R., Giraudel, J.L., Compin, A., 2001. Spatial analysis of stream invertebrates distribution in the Adour-Garonne drainage basin (France), using Kohonen self-organizing maps. Ecological Modelling 146, 167–180.

Chon, T.S., Park, Y.S., Moon, K.H., Cha, E.Y., 1996. Patternizing communities by using an artificial neural network. Ecological Modelling 90, 69–78.

Corne, S., Murray, T., Openshaw, S., See, L., Turton, I., 1999. Using computational intelligence techniques to model subglacial water systems. Journal of Geographical Systems 1, 37–60.

Darsono, S., Labadie, J.W., 2007. Neural-optimal control algorithm for real-time regulation of in-line storage in combined sewer systems. Environmental Modelling and Software 22, 1349–1361.

Dawson, C.W., Wilby, R.L., 2001. Hydrological modelling using artificial neural networks. Progress in Physical Geography 25 (1), 80–108.

Furundzic, D., 1998. Application example of neural networks for time series analysis: rainfall–runoff modelling. Signal Processing 64, 383–396.

Giraudel, J.L., Lek, S., 2001. A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination. Ecological Modelling 146, 329–339.

Hall, M.J., Minns, A.W., 1999. The classification of hydrologically homogeneous regions. Hydrological Sciences Journal 44 (5), 693–704.

Haykin, S., 1999. Neural Networks: A Comprehensive Foundation. Prentice Hall Upper Saddle River, New Jersey.

Hong, Y., Hsu, K., Sorooshian, S., Gao, X., 2005. Self-organizing nonlinear output (SONO): a neural network suitable for cloud patch-based rainfall estimation at small scales. Water Resources Research 41, W03008, doi:10.1029/2004WR003142.

Hsu, K., Gupta, H.V., Gao, X., Sorooshian, S., Imam, B., 2002. Self-organizing linear output map (SOLO): an artificial neural network suitable for hydrologic modeling and analysis. Water Resources Research 38 (12), 1302, doi:10.1029/2001WR000795.

Iliadis, L.S., Maris, F., 2007. An artificial neural network model for mountainous water-resources management: the case of Cyprus mountainous watersheds. Environmental Modelling and Software 22, 1066–1072.

Jain, A., Srinivasulu, S., 2006. Integrated approach to model decomposed flow hydrograph using artificial neural network and conceptual techniques. Journal of Hydrology 317, 291–306.

Kalteh, A.M., Berndtsson, R., 2007. Interpolating monthly precipitation by self-organizing map (SOM) and multilayer perceptron (MLP). Hydrological Sciences Journal 52 (2), 305−317.

Kohonen, T., 1982a. Analysis of a simple self-organizing process. Biological Cybernetics 44, 135−140.

Kohonen, T., 1982b. Self-organized formation of topologically correct feature maps. Biological Cybernetics 43, 59−69.

Kohonen, T., 2001. Self-Organizing Maps. Springer-Verlag, Berlin.

Kothari, R., Islam, S., 1999. Spatial characterization of remotely sensed soil moisture data using self-organizing feature maps. IEEE Transactions on Geoscience and Remote Sensing 37 (2), 1162−1165.

Lin, G., Chen, L., 2006. Identification of homogenous regions for regional frequency analysis using the self-organizing map. Journal of Hydrology 324, 1−9.

Maier, H.R., Dandy, G.C., 1996. The use of artificial neural networks for the prediction of water quality parameters. Water Resources Research 32 (4), 1013−1022.

Maier, H.R., Dandy, G.C., 1997. Determining inputs for neural network models of multivariate time series. Microcomputers in Civil Engineering 12 (5), 353−368.

Maier, H.R., Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environmental Modelling and Software 15, 101−124.

Moradkhani, H., Hsu, K., Gupta, H.V., Sorooshian, S., 2004. Improved streamflow forecasting using self-organizing radial basis function artificial neural networks. Journal of Hydrology 295, 246−262.

Murao, H., Nishikawa, I., Kitamura, S., Yamada, M., Xie, P., 1993. A hybrid neural network system for the rainfall estimation using satellite imagery. In: Proceedings of International Joint Conference on Neural Networks. IEEE press, pp. 1211−1214.

Obach, M., Wagner, R., Werner, H., Schmidt, H., 2001. Modelling population dynamics of aquatic insects with artificial neural networks. Ecological Modelling 146, 207−217.

Parasuraman, K., Elshorbagy, A., Carey, S.K., 2006. Spiking modular neural networks: a neural network modeling approach for hydrological processes. Water Resources Research 42, W05412, doi:10.1029/2005WR004317.

Raduly, B., Gernaey, K.V., Capodaglio, A.G., Mikkelsen, P.S., Henze, M., 2007. Artificial neural networks for rapid WWTP performance evaluation: methodology and case study. Environmental Modelling and Software 22, 1208−1216.

Schütze, N., Schmitz, G.H., Petersohn, U., 2005. Self-organizing maps with multiple input−output option for modeling the Richards equation and its inverse solution. Water Resources Research 41, W03022, doi:10.1029/2004WR003630.

Shanmuganathan, S., Sallis, P., Buckeridge, J., 2006. Self-organising map methods in integrated modelling of environmental and economic systems. Environmental Modelling and Software 21, 1247−1256.

Tran, L.T., Knight, C.G., O'Neill, R.V., Smith, E.R., O'Connell, M., 2003. Self-organizing maps for integrated environmental assessment of the Mid-Atlantic region. Environmental Management 31 (6), 822−835.

Vesanto, J., 1999. SOM-based data visualization methods. Intelligent Data Analysis 3, 111−126.

Vesanto, J., 2002. Data exploration process based on the self-organizing map. Ph.D. thesis. <http://lib.tkk.fi/Diss/2002/isbn9512258978/isbn9512258978.pdf> (accessed 09.02.06).

Vesanto, J., Alhoniemi, E., 2000. Clustering of the self-organizing map. IEEE Transactions on Neural Networks 11 (3), 586−600.

Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J., 2000. SOM Toolbox forMatlab5.<http://www.cis.hut.fi/projects/somtoolbox/package/papers/techrep.pdf> (accessed 09.02.06).

Wu, S., Chow, T.W.S., 2004. Clustering of the self-organizing map using a clustering validity index based on inter-cluster and intra-cluster density. Pattern Recognition 37, 175−188.