ANN-based groundwater level forecasting employing SOM clustering approach

VAHID NOURANI, FARNAZ DANESHVAR VOUSOUGHI, AIDA HOSSEINI BAGHANAM, MOHAMMAD TAGHI ALAMI

Abstract: The aim of this study is to present a mathematical based model to estimate the groundwater level (GWL) in Ardabil located at northwest of Iran, with association of some hydrological data (e.g., rainfall, discharge, etc.). In this way identifying various zones with similar groundwater level can be a promising idea which leads to appropriate overview on water table of the study area as well as efficient modeling. For this purpose, the Self Organizing Map (SOM) was used to cluster the homogenous monitoring piezometers in the plain by utilizing GWL and UTM (Universal Transverse Mercator) data. The sensitivity analysis was performed over normalized and non-normalized data of GWL and UTM in order to investigate their effects on clustering. Conventional K-Means method was applied to verify the results of SOM method. The central piezometer of each cluster was selected as a representative by means of statistical technique. Afterwards the three layer feed forward neural network (FFNN) model was utilized to calibrate a model via historical groundwater level records from the representative wells and relevant hydro-meteorological data. The last step was performed by simulating water table level of the representative piezometer from each zone of the plain via proposed model, to compare the computed and observed data. The results identify the specific piezometers that the groundwater level of them can represent the groundwater level in a particular region, thus adequate measures should be devoted on preserving such important monitoring piezometers. The modeling results can be utilized to frame the corresponding strategies to reduce the monitoring cost and to enhance the cost-effective benefits. The proposed methodology can be referred as a management plan for groundwater resources.

Keywords: Groundwater level; artificial neural network; clustering; self organizing map; Ardabil Plain

1 Introduction

For an effective management of groundwater, it is important to predict groundwater level fluctuations. Groundwater systems possess features such as complexity, nonlinearity, being multi-scale and random, all governed by natural and/or anthropogenic factors, which complicate the dynamic predictions. Therefore, many mathematical models have been developed to simulate this complex process [1]. Due to the recognized limitations of physically based models, data-driven methods such as artificial neural network (ANN) (as a black box models) has been widely used for forecasting in many areas of science and engineering. There are vast surge of reports on application of ANN in groundwater related field such as groundwater level-prediction or contamination.

Lallahem et al. used ANN methodology for estimating the groundwater level (GWL) in some piezometers implanted in unconfined chalky aquifer of Northern France. The reasonably good ANN-based simulations revealed the merit of using ANNs and specifically Multi Layer Perceptron (MLP) models [2]. Daliakopoulos et al. forecasted GWL by using ANNs [3]. They identified an optimal ANN architecture trained with a standard feed-forward neural network trained with the Levenberg–Marquardt algorithm that could simulate the decreasing trend of the groundwater level and provide acceptable predictions up to 18 months ahead. Nourani et al. evaluated the feasibility of ANN methodology in estimating the groundwater levels at some piezometers placed in an aquifer in north-western Iran [4]. They approved that ANNs based on the nonlinear intrinsic approach, provide accurate predictions. In another study, Yang et al. employed the Integrated Time Series (ITS) and Back-Propagation Artificial Neural Network (BPANN) models to forecast GWLs [5]. The results revealed the suitability of BPANN model in forecasting the GWL. Nourani et al. studied the spatiotemporal variations of the water level for the management of groundwater in the coastal areas [1]. They employed a hybrid, artificial neural network-
geostatistics methodology for spatiotemporal prediction of GWLs. Amir and Navid Jalalkamali employed a hybrid model of ANN and genetic algorithm (GA) in forecasting GWL in an individual well [6]. The consequences of their research admitted the superiority of ANN-GA model in prediction of GWL. Moreover, a method coupled of discrete wavelet transforms (WA) and ANN was proposed by Adamowski and Chan to predict GWL in Canada. The results of their study indicated the potential of WA-ANN models in forecasting GWL [7].

Cluster analysis is a statistical method of partitioning samples into homogeneous classes to produce an operational overview on data sets. The importance of cluster analysis techniques is evident in classifying multivariate data into sub-groups. There are diverse kinds of mathematical based methods in order to cluster data. The requirement to clustering techniques in hydrological and environmental filed especially groundwater related researches have been raised in recent decade. For instance, Kim et al. enjoyed the cluster analysis results in determination of hydro-geochemical characteristics of shallow groundwater in coastal area [8]. In a research, multivariate statistical techniques, such as cluster analysis and principal component analysis were performed on 22 well representative groundwater samples. The collected samples were analyzed for water quality variables [9]. Implanted piezometers all over the watershed can prepare informative data about the fluctuations of ground water level in a region. Thus the identification of dominant piezometers in a watershed that exhibit reliable data plays a crucial role in sufficient groundwater management. Application of clustering techniques on available piezometers interspersed spatially over the watershed can be an appropriate method to capture the adequate information of homogeneous piezometers. Indeed, by performing clustering techniques a spatial pre-processing approach is imposed on several piezometry data.

Among the various clustering methods, the conventional clustering methods (e.g., K-Mean) require the number of clusters be specified in advance, and their results are relevant to linear characteristics [10]. Unlike conventional methods, Self Organizing Map (SOM) as an un-supervised ANN, is used to cluster the complex data sets [11 and 12]. The SOM operates as an effective tool to convert complex, non-linear, statistical relationship between high-dimensional data items into simple, geometric relationship on a low-dimensional display so as to allow the number of clusters to be determined by inspection [13]. The SOM based classification is attractive, due to its topology preserving properties for solving various problems that traditionally have been the domain of conventional statistical and operational research techniques. SOMs have been successfully accepted widely in science and engineering problems; not only are their results unbiased, but they can also be visualized.

Peeters et al. compared the performance of the GEO3DSOM (a variant of the SOM-algorithm which is capable of explicitly incorporating three-dimensional spatial knowledge into the algorithm) and standard SOM in analyzing an artificial data set and a hydro-chemical data set [14]. Both techniques succeed very well in providing more insight in the groundwater quality data set, visualizing the relationships between variables, highlighting the main differences between groups of samples and pointing out anomalous wells and well screens. Chen et al. used the combination of the back-propagation network (BPN) and the self-organizing map (SOM) for forecasting the GWL data in southern Taiwan. According to their results, the multisite SOM-BPN model has the highest accuracy [15]. Kalteh et al. prepared a review paper to discuss the analysis, modeling and application of SOM approach in water resources [16], another review of SOM applications in meteorology and oceanography is reported by Liu and Weisberg as well [17]. The objective of this study is first clustering and next predict GWL pattern of the prominent piezometers at time scale of one month ahead in Ardabil Plain via SOM-ANN model.

2 Artificial Neural Network (ANN)

ANN is widely applied in hydrology and water resource studies as a forecasting tool. In ANN, feed-forward back-propagation (BP) network models are common to engineers. It has proved that BP network model with three–layer is satisfied for the forecasting and simulating in any engineering problem [18, 19]. As shown in Fig.1, three–layered feed forward neural networks (FFNNs), which have been usually used in forecasting hydrologic time series, provide a general framework for representing nonlinear functional mapping between a set of input and output variables. The term “feed-forward” means that a neuron connection only exists from a neuron in the input layer to other neurons in the hidden layer or from a neuron in the hidden layer to neurons in the output layer and the neurons within a layer are not interconnected to each other. The explicit expression for an output value of a three-layered FFNN is given by [11]:

\[
\hat{y}_k = f_o \left[ \sum_{j=1}^{N} w_{kj} f_h \left( \sum_{i=1}^{M} w_{ji} x_i + w_{j0} \right) + w_{ko} \right]
\]  
(1)
where $w_{ji}$ is a weight in the hidden layer connecting the $i$th neuron in the input layer and the $j$th neuron in the hidden layer, $w_{j0}$ is the bias for the $j$th hidden neuron, $f_a$ is the activation function of the hidden neuron, $w_o$ is a weight in the output layer connecting the $j$th neuron in the hidden layer and the $k$th neuron in the output layer, $w_{ko}$ is the bias for the $k$th output neuron, $f_o$ is the activation function for the output neuron, $x_i$ is $i$th input variable for input layer and $y_k$, $y$ are computed and observed output variables, respectively. $N_i$ and $M_k$ are the number of the neurons in the input and hidden layers, respectively. The weights are different in the hidden and output layers, and their values can be changed during the process of the network training.

3 Self organizing map

The self-organizing map (SOM) is an effective software tool for the visualization of high-dimensional data. Thereby it is able to convert complex, nonlinear statistical relationships between high-dimensional data items into simple geometric relationships on a low-dimensional display while preserving the topology structure of the data [13]. The way SOMs go about reducing dimensions is by producing a map of usually 1 or 2 dimensions which plot the similarities of the data by grouping similar data items together. The SOM network generally consists of two layers, an input layer and a Kohonen layer. The input layer is fully connected to the Kohonen layer, which in most common applications is two-dimensional. A two-level SOM neural network is a promising approach to catch a preliminary overview on intricate data set. It augments the conventional SOM network with additional one-dimensional Kohonen layer in which each neuron is connected to neurons in the previous Kohonen layer. The schematic view of the two-level SOM network is shown in Fig.1.

The SOM is trained iteratively: Initially the weights are randomly assigned. When the n-dimensional input vector $x$ is sent through the network, the distance between the weight $w$ neurons of SOM and the inputs is computed. The most common criterion to compute the distance is Euclidean distance [13]:

$$
\|x - w\| = \sqrt{\sum_{i=1}^{N} (x_i - w_i)^2} \quad (2)
$$

The weight with the closest match to the presented input pattern is called winner neuron or Best Matching Unit (BMU). The BMU and its neighboring neurons are allowed to learn by changing the weights at each training iteration, in a manner to further reduce the distance between the weights and the input vector [13]:

$$
\mathbf{w}(t + 1) = \mathbf{w}(t) + \alpha(t)h_{lm}(x - \mathbf{w}(t)) \quad (3)
$$

Where $\alpha$ is the learning rate, ranging in $[0 1]$, $l$ and $m$ are the positions of the winning neuron and its neighboring output nodes and $h_{lm}$ is the neighborhood function. The most commonly used neighborhood function is the Gaussian [13]:

$$
h_{lm} = \exp \left(-\frac{||l-m||^2}{2\sigma(t)^2}\right) \quad (4)
$$

where $h_{lm}$ is the neighborhood function of the best matching neuron $l$ at iteration $t$; and $l - m$ is the distance between neurons $l$ and $m$ on the map grid; and $\sigma$ is the width of the topological neighborhood. The training steps are repeated until convergence. After the SOM network is constructed, the homogeneous regions i.e., clusters, is defined on the map.

To evaluate the performance of clustering results produced by the SOM neural network, the silhouette coefficients are used as the measure of cluster validity [10]. The Silhouette coefficient of a cluster can indicate the degree of similarity of stations within a cluster, which is defined as [10]:

$$
S(i) = \frac{b(i) - a(i)}{\max(a(i),b(i))} \quad (5)
$$

where $S(i)$ is the silhouette of piezometer i, $a(i)$ measured as a Euclidean distance, is the average dissimilarity of cluster i to all other piezometers in cluster A; and $b(i)$ is the least average dissimilarity of piezometer i to the piezometers within a cluster different from cluster A. Thus, a smaller $S(i)$ value indicates a better similarity among piezometers within the same cluster. The overall quality of a clustering distribution can then be measured using the average silhouette width for the entire data set[10].

4 Study Region and Data

4.1 Study Region

Ardabil Plain (38°03’-38°27’ N and 47°55’-48°20’ E) located in the north-western Iran, covers an area of about 990 km² (see Fig.2). The mean areal annual rainfall over the Ardebil plain is about 304 mm.
May and August are known as the wettest and driest months of the region, respectively. The mean temperature in the Ardabil plain is about 9°C, and this plain is well known as the coldest region of Iran. The average number of freezing days in the Ardebil plain is about 130 days in a year. Fig. 3 presents the locations of piezometers in the study area. Over the plain, 17 piezometer are operated to measure groundwater level (P1, P2, P3, ..., P17). The data sampling has been reported in one-month interval at all of the pizeometers. The Plain is equipped with one stream gauge at the outlet and 6 rain gauges within the watershed, the average rainfall data have been used in this research. The monthly rainfall, runoff and groundwater level data are available for the water years of 22 years from 1988 to 2011.

4.2 Model Precision Evaluation

Input and output variables for ANN-based modeling purpose are usually normalized by scaling between zero and one to eliminate their dimensions. The following simple linear mapping of the variables is the most common method for this purpose [20].

\[
\tau_i = \frac{l_i - l_{\min}}{l_{\max} - l_{\min}}
\]

where, \(l_i\) is the actual value and is the respective normalized value. \(l_{\min}\) and \(l_{\max}\) are the minimum and maximum of the values, respectively.

The normalized data were divided into training and verification sets. In this study, two different criteria are used to measure the efficiency of the proposed methodology; the Root Mean Square Error (\(RMSE\)), and the determination coefficient (\(DC\)) [1]. The \(RMSE\) and \(DC\) demonstrate discrepancies between predictions and observations.

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{n} (l_i - \hat{l}_i)^2}{n}}
\]

\[
DC = 1 - \frac{\sum_{i=1}^{n} (l_i - \hat{l}_i)^2}{\sum_{i=1}^{n} (l_i - \bar{l}_i)^2}
\]

In Equations (7) and (8), \(n\) is the data number, \(l_i\) and \(\hat{l}_i\) are the observed data and the calculated values, respectively, and \(\bar{l}_i\) is the averaged value of the observed data. In the best model, \(DC\) and \(RMSE\) go to one and zero, respectively.

5 Results and Discussion

Due to the existence of various piezometers over the Ardabil Plain and the importance of managing groundwater resources, it is a necessity to unite the adequate information about the GWL in various regions of plain and identify the representative piezometers to predict GWL in future. Identification of representative piezometers can be useful while there are deficiencies at other piezometers over the plain. By combining the Self Organizing Map (SOM) as a clustering method to identify similar and dominant piezometers and ANN as a tool to capture the nonlinear underlying pattern involved at intricate data sets, a hybrid model of spatiotemporal GWL forecasting has been proposed in current study. The proposed model contained two stages.
rion was utilized to select the centroid piezometer of each cluster which is the best representative of the GWL pattern within the cluster. Sensitivity analysis was performed over the various SOM clusters with divers data sets as inputs. Firstly, the GWL data of 17 piezometers were applied as input data to the SOM. Secondly, the combination of GWL data and UTM (Universal Transverse Mercator) data of each piezometer were put to use in SOM as a multivariate clustering data. In both sorts of input variables the clustering approach were investigated by normalized and non-normalized data. The results in Table 1 revealed that the utilization of multivariate data in clustering did not lead to highly distinct results in comparing with the individual variable (solely GWL) data set. K-Means method was utilized in order to verify the SOM clustering results. To evaluate the performance of the clustering results produced by the SOM and K-Means, the Silhouette coefficient was used as the measure of cluster validity. The consequences led to selection of 1 by 3 Kohonen layer grid with normalized data of GWL. Subsequently the centroid piezometers were determined by Euclidean distance criterion. The central piezometers are the representative of GWL in various region of plain. Thus, the specific measures should be devoted on protection and maintaining such important piezometers in order to have precise and reliable data for predictions and management of plain. At the second stage, an ANN was trained for each of the representative piezometer to predict the GWL one month ahead. The model operates based on input data consist of rainfall and runoff amount at present month and GWL values of previous month. It is noticeable that the reason of such input order lies on sensitivity analysis over the various inputs especially the GWL lags which handle seasonality and autoregressive characteristics. The considerable fact was that the model efficiency did not alter excessively in cases other than one month lag. Therefore, in order to attain the appropriate results the minimum numbers of input data were selected to reduce the error possibility to minimum. The efficiency of the trained models was tested by the verification data. The results are depicted in Table 2.

### 6. Concluding Remarks

In this study the SOM, FFNN approaches were combined in order to develop a hybrid black box model for the multivariate GWL simulation of Ardabil Plain located in North-West Iran. First, an unsupervised neural network technique (i.e., SOM) was applied to distinguish the similar piezometric data and subsequently the dominant piezometer in each cluster, which detect the best GWL pattern of the region. The proper piezometer selection of SOM led to determination of regions with similar GWL in plain and also had this advantage that the groundwater related managements can be done sufficiently in future because of the data which are reported in good quality, the reason lies on a fact that by selection of permanent piezometers more equipments and investments on dominant piezometers can be devoted in order to have good performance in recording data. The FFNN GWL prediction model was constructed to find the non-linear relationship among the lagged GWL data with rainfall-runoff data. Overall, the results provide promising evidence in comparing the observed and predicted GWL. In order to complete the current study, it is recommended to use the presented methodology to forecast the GWL by adding other hydrological time series and variables (e.g., temperature or/and evapotranspiration) to the input layer of the model or augmentation of land use effects on the GWL at adjacent piezometers. Moreover, according to the uncertainty of the rainfall process and the ability of the Fuzzy concept in handling uncertainties, the conjunction of the ANN and FIS (Fuzzy Inference System) models, as an ANFIS (Adaptive Neural-Fuzzy Inference system) model could be a reliable choice for the model progress.

### Table 1. SOM clustering results.

<table>
<thead>
<tr>
<th>Applied data</th>
<th>SOM Clustered Piezometers</th>
<th>Silhouette Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>GWL</td>
<td>2, 3, 17</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>1, 4, 5, 6, 9, 16</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>7, 8</td>
<td>0.87</td>
</tr>
<tr>
<td>Normalized</td>
<td></td>
<td>0.87</td>
</tr>
<tr>
<td>GWL &amp; UTM</td>
<td>2, 3, 17</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>1, 4, 5, 6, 9, 16</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>7, 8</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.87</td>
</tr>
<tr>
<td>Normalized</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GWL</td>
<td>1, 4, 5, 6, 9, 16</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>7, 8</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.87</td>
</tr>
<tr>
<td>Normalized</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GWL</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 2. Results and structures of FFNN model

<table>
<thead>
<tr>
<th>#Piezometer</th>
<th>#Network structure</th>
<th>DC Train</th>
<th>DC Verify</th>
<th>RMSE Train</th>
<th>RMSE Verify</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2</td>
<td>3-3-1</td>
<td>0.77</td>
<td>0.68</td>
<td>0.00027</td>
<td>0.00011</td>
</tr>
<tr>
<td>P5</td>
<td>3-9-1</td>
<td>0.97</td>
<td>0.62</td>
<td>0.00055</td>
<td>0.00056</td>
</tr>
<tr>
<td>P7</td>
<td>3-5-1</td>
<td>0.95</td>
<td>0.86</td>
<td>0.00033</td>
<td>0.00032</td>
</tr>
</tbody>
</table>

* COMB: G(t-1): GWL with 1-month lag, I(t): Precipitation, Q(t): Runoff
* first, second and third number present input variable, hide neurons and output variable, respectively.
Acknowledgments: The first author of the paper has been invited as plenary speaker of the conference. The kind invitation of the WSEAS is appreciated.

Reference