MIDDLE POMERANIAN SCIENTIFIC SOCIETY OF THE ENVIRONMENT PROTECTION

Rocznik Ochrona Środowiska

Volume 21. Year 2019 ISSN 1506-218X 29-51

The Investigation of the Applicability of Data-Driven Techniques in Hydrological Modeling: The Case of Seyhan Basin

Evren Turhan, Mümine Kaya Keleş,*
Atakan Tantekin, Abdullah Emre Keleş
Adana Alparslan Türkeş Science and Technology University, Turkey
**corresponding author's e-mail: eturhan@atu.edu.tr*

1. Introduction

Hydrology science is described as the life cycle of water. It is a fact that rainfall-runoff modeling and the other data-driven techniques are significant events in this cycle. Especially, estimation of missing the streamflow is an important process for designing and planning flood protection studies in the data-driven techniques. Drought is thought one of harmful natural phenomenon in terms of hydrology. Drought estimation methods should be investigated in detail. Furthermore, observational data in hydrological works have many errors due to gauging. These obtained data should be evaluated for applicability in some hydrological calculations. Therefore, this paper aims to research use of data-driven techniques in hydrology.

In the literature there are several studies that use ANN and data mining methods for hydrological studies such as rainfall-runoff modeling and drought analysis. The ANN methods is one of the most commonly used in rainfall-runoff relation modeling (Dawson & Wilby 1999, Alp & Cigizoglu 2005, Nourani et al. 2009, Machado et al. 2011, Gumus et al. 2013, Kumar et al. 2016, Turhan et al. 2016a, Loyeh & Jamnani 2017, Patel & Joshi 2017, Asadi et al. 2019, Lin et al. 2019). Feed Forward Back Propagation (FFBPNN) and Generalized Regression Neural Networks (GRNN) methods can be seen to be generally utilized for hydrological data estimation (Cigizoglu 2005, Turan & Yurdusev 2009, Turhan et al. 2016b, Tayyab et al. 2016, Gumus et al. 2018). Standardized Precipitation Index (SPI), Standardized Runoff Index (SRI), Streamflow Drought Index (SDI), De Martonne Index (DMI),...etc methods play a very important role in determining drought (Shukla & Wood 2008, Stachowski 2010, Sattari et al. 2011, Bartholy et al. 2013, Tabari et al. 2013, Turhan et al. 2016c, Gumus & Algin 2017, Gumus 2017, Myronidis et al. 2018, Tri et al. 2019). The studies conducted in the field of construction/civil engineering using data mining methods have been increasing in recent years. It is observed that the results are taken by using data mining methods in the studies including concrete compressive strength, leadership analysis, productivity determination, building material decision making process, rainfall estimation, streamflow estimation, occupational health and safety assessment, cost analysis, analysis of traffic accidents, project management,... etc. subjects (Caldas et al. 2002, Wilmot & Cheng 2003, Baykasoglu 2005, Liaoa & Perng 2008, Kaya et al. 2013, Keles & Kaya 2014, Ozel & Topsakal 2014, Keles 2016, Kaya Keles 2017, Kaya Keles & Keles 2017). Data mining techniques have been used for hydrological studies (Trafalis et al. 2002, Damle & Yalcin 2007, Terzi 2012, Yurekli et al. 2012, Kusiak et al. 2013, Keskin et al. 2013, Sattari et al. 2018, Hatami et al. 2018, Mishra et al. 2018, Sattari & Sureh 2019, Kaur & Sood 2019, Sezen et al. 2019).

In this paper, two data-driven techniques such as ANN and Data Mining were investigated in terms of availability in hydrology works. FFBPNN and GRNN methods were examined on the rainfall-runoff modeling for ANN. ANN algorithm was created using Matlab software. Besides, hydrological drought analysis were examined using data mining techniques. Drought analysis was carried out using the SRI method. Seyhan Basin was preferred to carry out these techniques. As compared to many data mining works with hydrological elements, few works are present in the literature, thus using data mining techniques for streamflow data can be a novelty of this study. There is a need for new studies that will enable the use of data mining methods in the field of hydrology in order to fill this gap in the literature. Due to the lacking number of available hydrological and meteorological data, it is necessary to complete missing data in the basin modeling/hydraulic structures design studies and estimate for the future. Examining these data in the same basin could provide great contributions in order to investigate applicability. Consequently, it is thought that the application of these different techniques (ANN and data mining) in the same basin could make a great contribute to the literature. Also, the use of several hybrid data mining methods can be planned in future studies.

2. Materials and Methods

2.1. Artificial Neural Networks (ANN) Methods

2.1.1. Feed Forward Back Propagation (FFBPNN) Method

This method is commonly used in ANN studies. The method consists of input, hidden and output layers. The output of cells in a layer provides input values, by means of weights, to the next layer (Dawidowicz et al. 2018). As the input layer processes the data obtained with the help of a weighted coefficient input vector, this input layer transmits along the cell structures in the hidden layer. Therefore, the output data is produced by changing the hidden and output layers. Using a back propagation algorithm, it tries to provide good convergence (Turhan et al. 2016a).

In this method:
m – the layer number,
${X_i}^m$ – the input data of *i* unit in *m*th layer,
${y_i}^m$ – the output data of *i* unit in *m*th layer,
${w_{ij}}^m$ – weight coefficient linking *i* unit in (*m-1*)th layer to *j* unit in *m*th layer,
w – a random real value in the additional forward values calculated for each *j* units in *m*th layer.

Then, the output data are obtained from Equation 1 (Gumus & Kavsut 2013):

$$ {y_i}^m = f(\textstyle\sum_i {y_i}^{m-1}\, {w_{ij}}^m) \tag{1}$$

For the output, the error terms shown as δ are calculated by Equation 2:

$$ {\delta_i}^m = \left({y_i}^M - {y_i}^m\right) f'\left({X_i}^M\right) \tag{2}$$

Also, the error terms are calculated for the backward hidden layer units by Equation 3:

$$ {\delta_i}^{m-1} = f'\left({X_i}^{m-1}\right) \textstyle\sum_i {\delta_i}^m\, {w_{ij}}^m \tag{3}$$

All these weights are obtained by Equation 4:

$$ {w_{ij}}^{next} = {w_{ij}}^{previous} + \Delta {w_{ij}}^m \tag{4}$$

η is the learning coefficient and changes in the weight coefficient during learning are shown by Equation 5:

$$ \Delta {w_{ij}}^m = \eta {\delta_i}^m {y_i}^{m-1} \tag{5}$$

The process is repeated for each step until the total error reaches a minimum value (Konate et al. 2015).

2.1.2. Generalized Regression Neural Networks (GRNN) Method

Regression of dependent variable *y* by independent variable *x* is shown by Equation 6 (Kisi 2006):

$$E\left[\frac{y}{x}\right] = \frac{\int_{-\infty}^{\infty} y\, f(x,y)dy}{\int_{-\infty}^{\infty} f(x,y)dy} \tag{6}$$

If the probability function is not known, this function can be estimated from X_i and Y_i values by Equation 7:

$$f(X,Y) = \frac{1}{(2\pi)^{(p+1)/2}\, s^{(p+1)}}\frac{1}{n}\sum_{i=1}^{n} \exp\left[-\frac{(X-Xi)^T(X-Xi)}{2s^2}\right]\exp\left[-\frac{(Y-Yi)^2}{2s^2}\right] \tag{7}$$

where:
p – the size of the x vector,
n – the number of observed data,
s – the correction parameter.

D_i^2 is to be regarded as a scalar function in Equation 8:

$$D_i^{\,2} = (X - X_i)^T(X - X_i) \tag{8}$$

Consequently, the dependent variable is obtained by Equation 9:

$$Y(X) = \frac{\sum_{i=1}^{n} Y_i \exp(-\frac{D_i^{\,2}}{2s^2})}{\sum_{i=1}^{n} \exp(-\frac{D_i^{\,2}}{2s^2})} \tag{9}$$

Figure 1 shows the general schematic structure of the FFBPNN and GRNN.

The formula used in the normalization process can is shown below (Equation 10):

$$Q_n = \psi\, \frac{Q_e - Q_{min}}{Q_{max} - Q_{min}} + \rho \tag{10}$$

Q_{max} and Q_{min} are the maximum and minimum flow values for each Flow Observation Station (FOS), respectively. Q_e indicates the normalized flow value, Ψ and ρ are evaluated to be constants the values of *0.6* and *0.2*, respectively. *N* represents total data, normalization has been completed by performing an inverse equalization process (Turhan et al. 2016a).

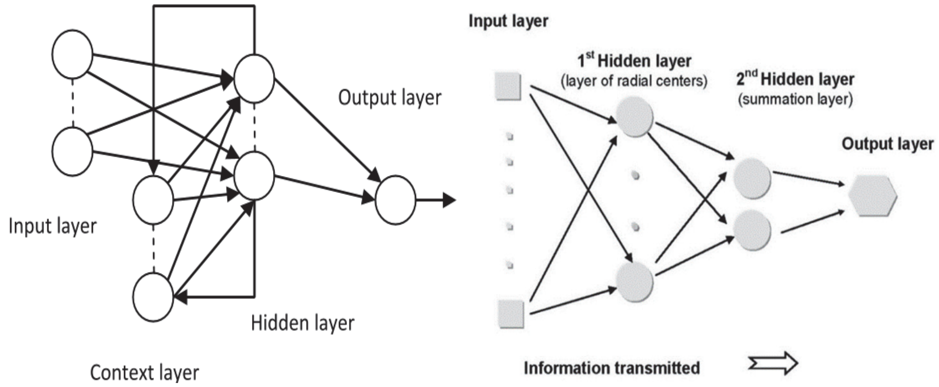

Fig. 1. General schematic structure of FFBPNN (left) and GRNN (right) (Ikiel & Ozyildirim 2013, Cigizoglu 2005)

Mean Square Error (MSE) and R^2 (Coefficient of determination-R^2) formulas are shown Equation 11 and 12:

$$\mathrm{MSE} = \frac{1}{\mathrm{N}} \tag{11}$$

$$\mathrm{R}^2 = \frac{\sum_{i=1}^{N}(Q_{observed}-Q_{average})^2-\sum_{i=1}^{N}(Q_{observed}-Q_{calculated})^2}{\sum_{i=1}^{N}(Q_{observed}-Q_{average})^2} \tag{12}$$

2.2. Drought analysis methods

2.2.1. Standardized Runoff Index method (SRI)

Different drought indices have been used for hydrological works in the literature. The Standardized Precipitation Index (SPI), Standardized Precipitation Evapotranspiration Index (SPEI), Evapotranspiration Deficit Index (ETDI) and Soil Moisture Deficit Index (SMDI) are based on precipitation, precipitation-temperature, evapotranspiration and soil moisture, respectively. In these indices, while calculating the SPI value, firstly the precipitation value at a certain time is subtracted from the average precipitation value. Then, this value is divided by the standard deviation and the SPI value is then obtained (Shukla & Wood 2008). Besides these, the SRI based on runoff or streamflow data, describes the events of hydrological drought. SRI method is largely similar to the SPI method except for the use of streamflow values instead of precipitation values. Formulas for the SRI are modified by setting the these streamflow data. The SRI value is calculated with the help of formula below (Equation 13):

$$\mathrm{SRI} = \frac{X_i - X_i^{mean}}{\sigma} \tag{13}$$

where:
X_i – the monthly average streamflow values over a certain time period,
X_i^{mean} – the average of monthly streamflow values within a selected time period,
σ – the standard deviation value.

Furthermore, the standard deviation value is calculated as shown below (Equation 14):

$$\sigma = \sqrt{\frac{n\sum x^2-(\sum x)^2}{n(n-1)}} \tag{14}$$

where:
n – the number of observations for a given month,
x – the mean streamflow value of that month.

Calculation of the index shows a complex structure due to the fact that the streamflow data doesn't conform to normal distribution during periods of *12* months or less. Accordingly, first of all, the streamflow series are made to conform to this distribution. Drought seasons can be determined for a selected time period by applying normalization procedures to the SRI values.

In the drought analysis using SRI values, the negative time period in which the index takes a continuous negative value is determined as the drought season (Kumar et al. 2009, Kubiak-Wójcicka & Bak 2018). The month when the index falls below zero is regarded as the starting point of the drought, and the month when the index has a positive value is predicted as the end point of the drought. The classification of the SRI value-drought categories, prepared using the SPI method, is shown in Table 1 (Keskin et al. 2007).

Table 1. Classification of the SRI Value – Drought Category (Keskin et al. 2007)

SPI Value	Drought Category	SPI Value	Drought Category
≥ 2	Extreme Rainfall	0.00-(-0.99)	Mild Drought
1.50-1.99	Severe Rainfall	-1.00-(-1.49)	Moderate Drought
1.00-1.49	Moderate Rainfall	-1.50-(-1.99)	Severe Drought
0.99-0.00	Mild Rainfall	≤ -2	Extreme Drought

2.3. Data mining methods

Data Mining is the processing of large amounts of data with automatic or semi-automatic methods to find meaningful patterns. This method is an interdisciplinary study. Statistics and machine learning are the scientific disciplines on which data mining is based most. The underlying meaning of data mining is the knowledge discovery in databases (KDD) process. KDD refers to the broad process of finding knowledge in data. The seven steps of the KDD is shown as in Figure 2 (Kaya Keles 2017). The KDD uses data mining methods to identify knowledge, according to the characterization of measures using preprocessing and transformations of the database.

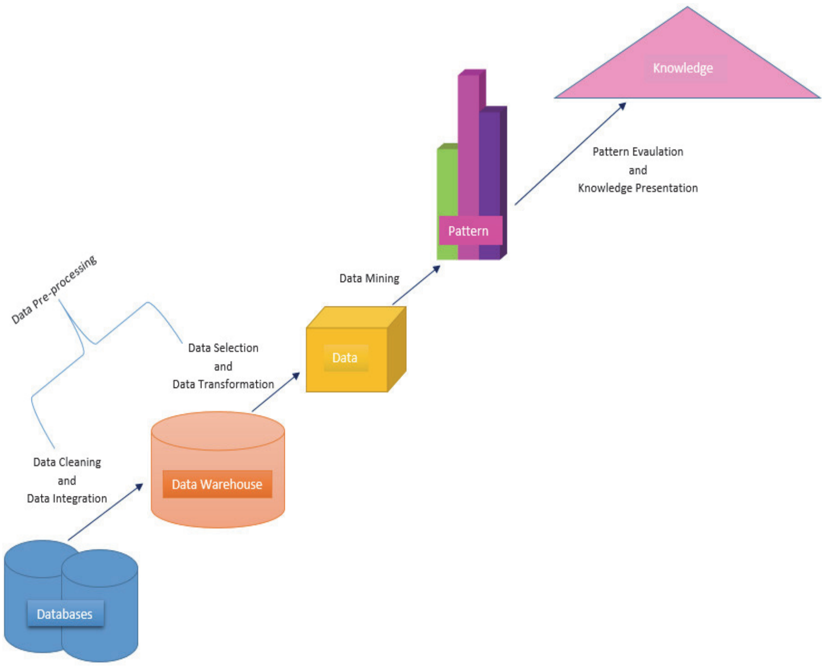

Fig. 2. Knowledge Discovery in data bases (KDD) Process (Kaya Keles 2017)

Waikato Environment for Knowledge Analysis (Weka) software was used for data mining process. The Weka is one of the packages used in data mining and machine learning which are the important subjects of computer science. It is a data analysis tool under the GNU (General Public License) developed by the University of Waikato with java language. It is basically a data mining program where machine learning algorithms and data pre-processing needs are combined together. It uses methods such as clustering, classification and association rule mining to perform data mining process. The file extension format is *'.arff'*. In the field of civil engineering where there are many sub-disciplines, due to the increase in knowledge in parallel with technological developments, information needs to be processable.

2.3.1. Multi-Layer Perceptron (MLP)

MLP called multi-layer feedforward network, was proposed as an enhancement of the perceptron model. This model is the most widely used neural network architecture for classification tasks. The main characteristics of the MLP are:

- The entire network is hierarchically trained. It uses back propagation learning.
- Sigmoidal units are used in hidden layers and in the output layer.
- It can make more modeling using at least 1 hidden layer (Aggarwal, 2014).

Although, in theory, MLPs are considered to be capable of performing any sort of classification task, in practice, in cases of insufficient number of hidden layers or insufficient training, data classification performance may be low. As shown in Figure 3, the units in the network are arranged in layers including an input layer, several hidden layers and an output layer.

The layers are organized in sequential order that follows the data flow from the input layer to the hidden layers and ends at the output layer. In this process known as forward propagation of the network, each unit in a layer receives data from the previous layer's units and the output units ensure the final decision of the classification. Forward propagation refers to the transfer of data from the input layer to the output layer, while the back propagation refers to the spread of error derivatives from the output layers to the hidden layers. The back propagation training algorithm as shown in Figure 4, which is designed to minimize an error criterion of the feed forward network, is an iterative gradient descend algorithm.

The MLP classifier in Weka is a classifier that uses back-propagation to learn a multi-layer perceptron when classifying samples. This classifier is located under the functions tab in Weka under the name MLP. In this study, the default values are preferred when using MLP. For learning rate, the value of *0.3* was used. For momentum rate, the value of *0.2* was used. For threshold, the value of *20* was used. With Weka, the user can find the classification accuracy by measuring the success rate with cross-validation method. In this paper, 10-fold cross validation method was used with MLP classifier.

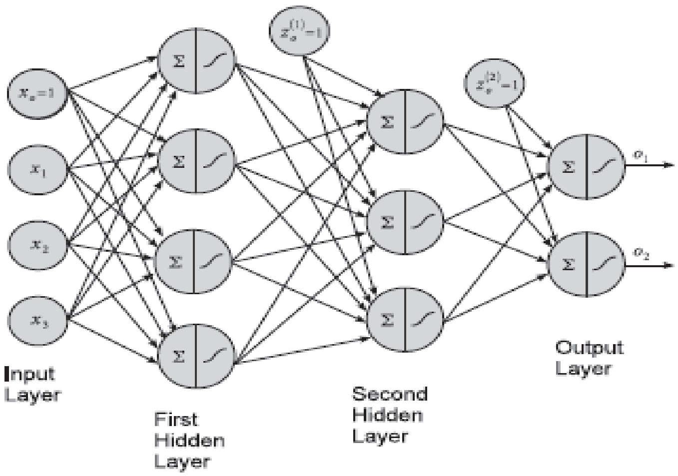

Fig. 3. An MLP network made of one input layer, two hidden layers and one output layer (Aggarwal, 2014)

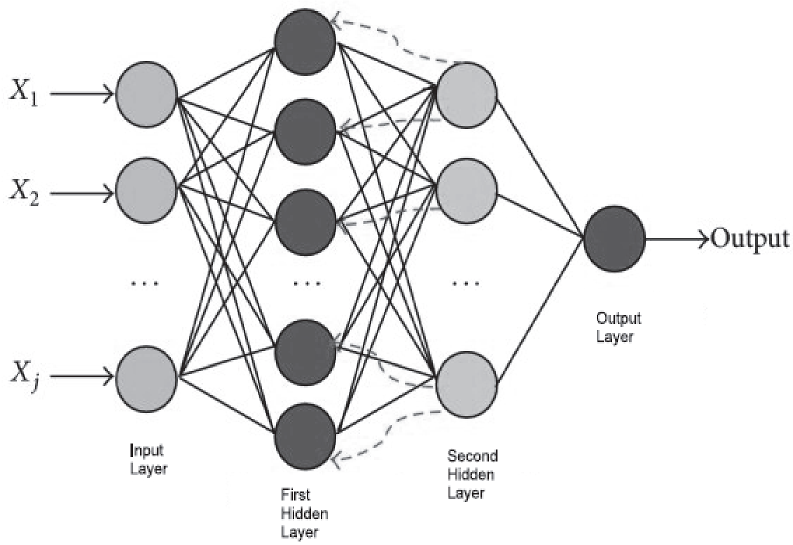

Fig. 4. An MLP network made of one input layer, two hidden layers and one output layer with backpropagation (Aggarwal, 2014)

2.4. Case study

All input data were obtained from the Seyhan Basin observed data for analysis. The Seyhan Basin is one of the most important basins in the south of Turkey in terms of social, economic, agricultural and numerous endemic species. The Seyhan Basin's area is approximately 20450 km^2 and covers of the 2.82% of Turkey with regard to surface area (Topcu & Seckin 2016). Thus, it is thought that the Seyhan Basin is an interesting basin for the hydrological studies (Fujihara et al. 2008, Tuncok 2016, Cavus and Aksoy 2019). The location of the Seyhan Basin can be seen in Figure 5. The average values for each FOS and Precipitation Observation Station analyzed, observed data and the time periods are given in Table 2.

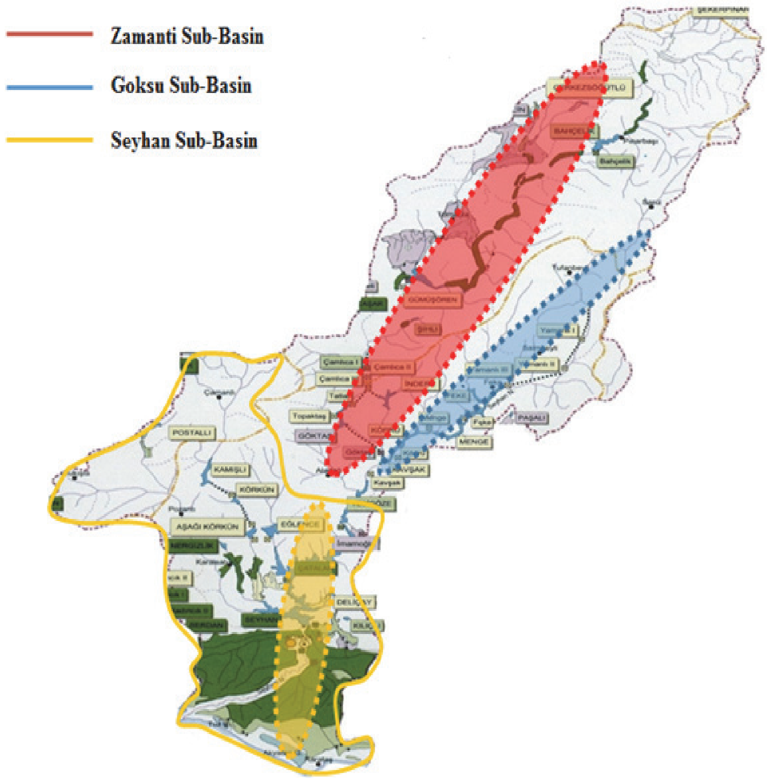

Fig. 5. A map of the Seyhan Basin (The Ministry of Forestry & Water Affairs (MFWA), 2016)

Table 2. Parameters of FOS and POS (Electrical Work Surveying Administration (EWSA), 2008)

FOS-POS	Location	Average Flow Values (m^3/s)	Average Precipitation Values (mm)	Observed Periods
1801 Göksu River FOS	Kozan-Saimbeyli	30	-	1936-2005
1818 Seyhan River FOS	Adana and on Uctepe Bridge	145	-	1966-2005
1821 Eglence Stream FOS	Adana-Karaisali	10.7	-	1971-2000
1822 Zamantı River FOS	Border of Adana-Kayseri	20	-	1970-2005
Catalan POS	Adana-Karaisali	-	819.54	1964-1987
Karaisalı POS	Adana-Karaisali	-	910.57	1957-2000

3. Results and discussion

3.1. Artificial Neural Networks (ANN) results

Monthly average flow and precipitation data in the Seyhan Basin were used as input data for the ANN methods. For both models, five data were provided as input data, and two intermediate layers were used (Figure 6). As a result, one output value was determined. All data were scaled from *0.1* to *0.9*, and a logarithmic sigmoid transfer function was used. The output layer was a linear function. For this reason, the data was normalized without entering the network structure. Even though meteorological data were very high for the ANN, normalization was easily performed. The algorithm required for calculation was created and simulated in Matlab software. Seventy percent (70%) of these data were available for training and the rest (30%) for testing. The MSE value was close to zero (0) and R^2 was close to the value of *1*, indicating a good result was predicted. The value of two neurons or the cell number was used in the hidden layer. The output value was the value of *1*. Many trials were performed and only the five models giving the best results are shown in Table 3.

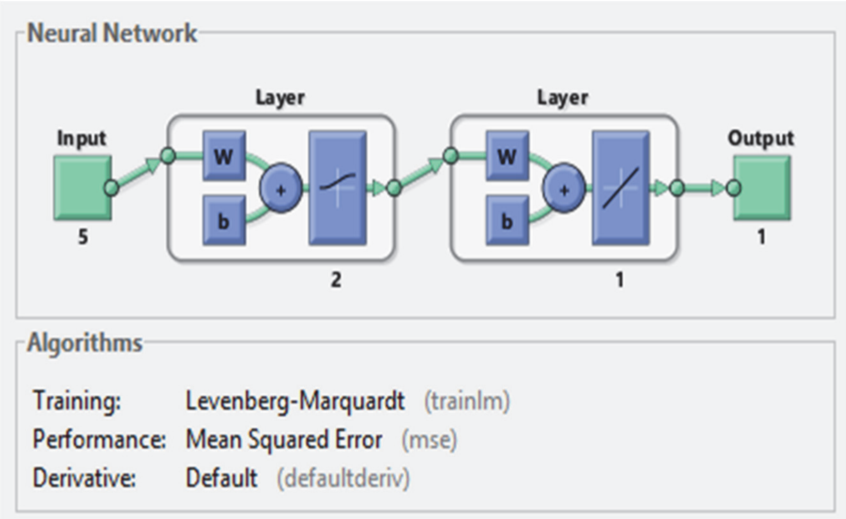

Fig. 6. ANN structure

Table 3. FFBPNN and GRNN Results for Training and Testing

Model No	FFBPNN				GRNN			
	Training		Testing		Training		Testing	
	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE
1	0.898	20.222	0.945	5.625	0.967	6.658	0.799	23.602
2	0.977	5.208	0.946	6.817	0.989	2.544	0.894	7.877
3	0.972	5.206	0.945	6.818	0.988	2.550	0.892	7.892
4	0.977	5.209	0.943	6.812	0.989	2.536	0.894	7.880
5	0.977	7.965	0.945	12.515	0.987	4.405	0.912	12.406

According to Table 3, the FFBPNN is the best model, giving the highest R^2 and lowest MSE values. The structure of model numbers can be shown as five input and one output data in the following:

- Model No-1: $Pcatalan_t$, $Pkaraisali_t$, $Q1801_t$, $Q1822_t$, $Q1818_{t+1}$--- $Q1818_t$
- Model No-2: $Pcatalan_t$, $Pkaraisali_t$, $Q1801_t$, $Q1818_t$, $Q1822_{t+1}$--- $Q1822_t$
- Model No-3: $Pcatalan_t$, $Pkaraisali_t$, $Q1818_t$, $Q1822_t$, $Q1801_{t+1}$--- $Q1801_t$
- Model No-4: $Pcatalan_t$, $Pkaraisali_t$, $Q1801_t$, $Q1822_t$, $Q1818_{t-1}$--- $Q1818_t$
- Model No-5: $Pcatalan_t$, $Pkaraisali_t$, $Q1801_t$, $Q1818_t$, $Q1822_{t-1}$--- $Q1822_t$

In Figure 7, for trial number-2 flow (m3/s)-time (month), the graphs were created and compared to the calculated and observed data. In Figure 8 for the same trial (no 2), the calculated flow data–observed data (m^3/s) were drawn.

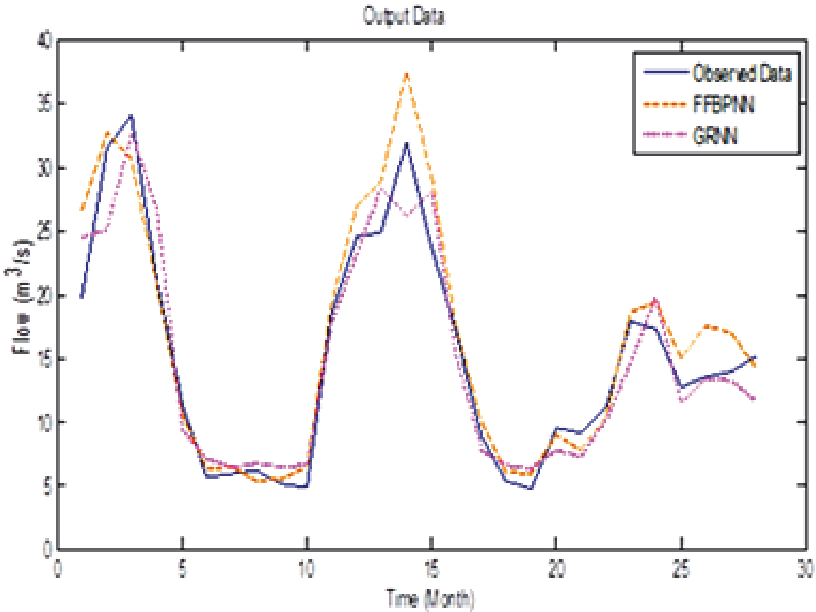

Fig. 7. Comparison between flows (m^3/s)-time (month)

When Figure 7 and 8 are examined in detail, the FFBPNN values were quite good results according to the observed data. Although the GRNN calculated a higher R^2 value than the FFBPNN in the training phase, FFBPNN calculated the highest R^2 and least MSE values in test phase**.** The best result of the FFBPNN model is the input data $Pcatalan_t$, $Pkaraisali_t$, $Q1801_t$, $Q1818_t$, $Q1822_{t+1}$ and the output data is $Q1822_t$ (structure 5-2-1-1). Many studies can be seen higher FFBPNN results than the GRNN (Gumus & Kavsut 2013, Tayyab et al. 2016). However, it can be seen that GRNN and other artificial neural network methods produce high values in network structure changes (Gumus et al. 2013, Turhan et al. 2016a).

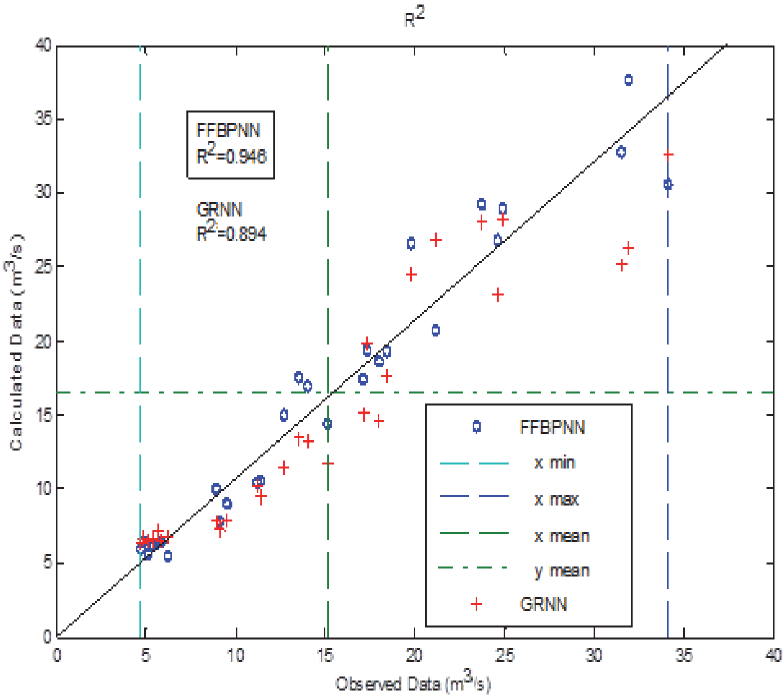

Fig. 8. Comparison between calculated and observed flow data (m^3/s)

3.2. Meteorological and hydrological drought analysis results

SRI drought analysis results for *1801, 1818, 1821* and *1822* can be seen in Figure 9. Drought analysis calculations were performed with the help of MS Excel and missing data were completed by using the ANN methods.

3.3. MLP prediction results

The generated SRI data, *1818*$_{t-1}$ data, *1818*$_{t-2}$ data, *1818*$_t$ monthly flow data for *1818* FOS were used to predict the class of the drought. To analyze drought, Table 4 (McKee et al. 1993, World Meteorological Organization (WMO) 2012), which is the SRI values table was used and data were normalized using this table for SRI analysis. Since data mining classification algorithms aim to increase the success of classification with fewer classes, SPI values in Table 4 consisting of 7 classes were used in this study.

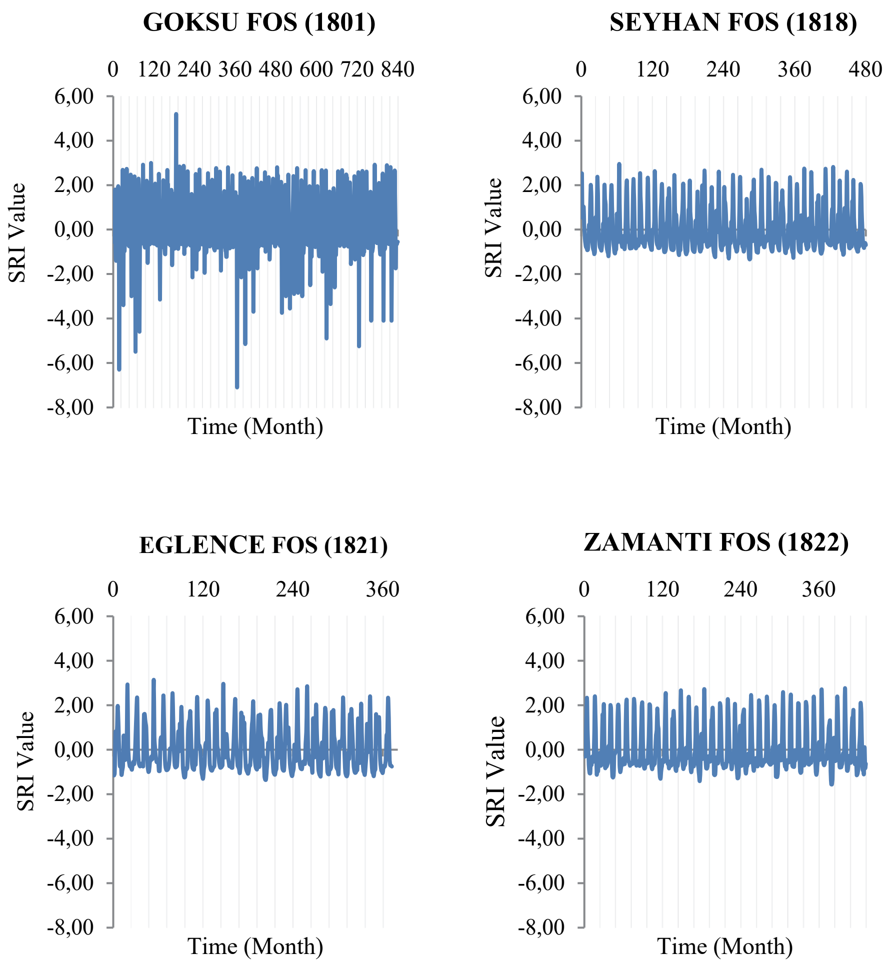

Fig. 9. Hydrological drought analysis results

Table 4. SRI values for Data Mining methods

2.0+	extremely wet
1.5 to 1.99	very wet
1.0 to 1.49	moderately wet
-.99 to .99	near normal
-1.0 to -1.49	moderately dry
-1.5 to -1.99	severely dry
-2 and less	extremely dry

This table was used to test and measure the success of the SRI value using the data mining method. According to Table 4, SRI values were classified as extremely wet, very wet, moderately wet, near normal, moderately dry, severely dry, and extremely dry. To measure the validation of the system which used data mining techniques, classification accuracy, F-Measure, Root Mean Squared Error methods were used. All processes were done for 1801, 1821 and 1822 FOS, too. Classification Accuracy shown in Equation 15 (Powers 2011) is the ratio of number of correct predictions to the total number of input samples.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Input Samples}} \quad (15)$$

F-Measure shown in Equation 16 (Powers 2011) is also called as F1 Score, F-Score, or F-Value. It is the weighted average or harmonic mean of precision and recall. It ranges from the values of *0* to *1*. Precision means positive predicted value, while recall means sensitivity.

$$\text{F} - \text{Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (16)$$

Root Mean Squared Error (RMSE) shown in Equation 17 (Kamble & Deshmukh 2019) is evaluated as the root squared value of the total of differences between the probability distribution output from the classification algorithm and the vector of probabilities symbolizing the real class of all samples (Mehdiyev et al. 2016). The RMSE is a quadratic scoring rule which sets the medium magnitude of the error (Guo et al. 2015).

In RMSE formula, $\sum$ means summation, x_i means predicted values, $\bar{x}_i$ means observed values, *N* means sample size.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x}_i)^2} \quad (17)$$

After the missing data was completed, the data set was analyzed with MLP algorithm to predict the drought types using Table 4. For each flow observation stations including *1818, 1821, 1801 and 1822*, all SRI values were calculated for 3, 6 and 12 months. The classification accuracy, F-Measure and the RMSE values of MLP algorithm is shown in Table 5 for each station and the SRI values.

Table 5. MLP results for different SRI time series

FOS	SRI values for 3, 6 and 12 months	MLP Results		
		Performance Evaluation Criteria		
		F-Measure	Accuracy	RMSE
1818	SRI-3	0.883	89.25%	0.157
	SRI-6	0.656	73.66%	0.234
	SRI-12	0.721	77.96%	0.218
1821	SRI-3	0.829	84.95%	0.182
	SRI-6	0.672	75.00%	0.229
	SRI-12	0.677	75.54%	0.224
1801	SRI-3	0.789	83.87%	0.229
	SRI-6	0.737	80.91%	0.210
	SRI-12	0.751	80.11%	0.199
1822	SRI-3	0.834	88.71%	0.170
	SRI-6	0.542	64.52%	0.256
	SRI-12	0.686	76.34%	0.222

According to this table, the success of MLP algorithm was greater than 65% for all stations and SRI values. But for 3-months SRI values, the success of MLP was greater than 84%. It shows that using 3-months SRI values, the drought type of the next period can be predicted with a higher success rate. It was observed that the success rate of MLP in all stations was higher when using SRI-3 month data. The estimation success of the 3-month data is followed by 12-month data. The least success was the 6-month SRI data. According to these results, MLP algorithm showed the highest success in drought estimation with 3-month data (SRI-3) as 89.25% accuracy rate and 0.883 F-Measure value for the *1818 FOS*. For *1821 FOS*, the MLP algorithm showed the highest success in drought estimation with 3-month data (SRI-3) as 84,95% accuracy rate and 0.829

F-Measure value. For *1801 FOS*, the MLP algorithm showed the highest success in drought estimation with 3-month data as 83,87% accuracy rate and 0.789 F-Measure value. For *1822 FOS*, the MLP algorithm showed the highest success in drought estimation with 3-month data as 88,71% accuracy rate and 0.834 F-Measure value. This shows that by using data mining methods, it is possible to estimate the drought type of the next period with a success of over 83% using SRI-3 month data.

When compared with other studies in the literature, it was observed that the results obtained within the scope of the study produced similar results with Sattari's study (2011). In both studies, it was observed that the estimation ability of the ANN model was weakened as the estimation time increased. When Terzi's study (2012) was compared with this study, it was observed that similar results were reached. In contrast to this study, although the MLR algorithm, which is a regression based algorithm, is used in Terzi's study, according to the results obtained, just like in this study, data mining methods can be used successfully in monthly estimates in the field of hydrology. Also, as in the results of MLP, which is a function-based algorithm obtained in this study, in Sattari's study (2017), Support Vector Regression, which is a function-based algorithm produced higher accuracy rate than tree-based M5 tree algorithm.

4. Conclusion

In this study, ANN methods such as Feed Forward Back Propagation (FFBPNN) and Generalized Regression Neural Networks (GRNN) were used for rainfall-runoff relationship modeling. An ANN algorithm was created using the software program. It is seen that the FFBPNN method is the best model for the ANN in terms of giving the highest R2 and lowest MSE values. Subsequently, hydrological drought analysis was carried out using the SRI analysis. MLP method, which is one of the data mining methods, was performed to predict the drought type according to limit values. The generated monthly flow data, for 1818, 1801, 1821 and 1822 FOS were used to predict the class of drought using the SRI tables. With these SRI values, data from the data set prepared within the scope of the study were classified and the type of drought was tried to be analyzed using MLP algorithm. MLP algorithm showed the highest success in drought estimation with 3-month data (SRI-3) as 89.25% accuracy rate and 0.883 F-Measure value for the 1818 FOS. The least success was the 6-month SRI data. Consequently, it is foreseen that these techniques may be used easily in terms of modeling of hydrological cycle elements for completion missing data or the other basin planning and management works.

References

Aggarwal, C.C. (2014). Data classification: algorithms and applications. *CRC Press*, 1-64.

Alp, M., & Cigizoglu, H. K. (2005). Modelling rainfall-runoff relation using different artificial neural network methods. *2th National Water Engineering Symposium*, 589-598, Izmir.

Asadi, H., Shahedi, K., Jarihani, B., & Sidle, R. C. (2019). Rainfall-runoff modeling using hydrological connectivity index and artificial neural network approach. *Water*, *11*(2), 1-20.

Bartholy, J., Pongracz, R., & Sabitz, J. (2013). Analysis of drought index trends for the Carpathian Basin using regional climate model simulations. *Geophysical Research Abstracts*, EGU General Assembly, Vienna, Austria, 15.

Bayissa, Y., Maskey, S., Tadesse T., Van Andel S.J., Moges S., Van Griensven, A., & Solomatine, D. (2018). Comparison of the Performance of Six Drought Indices in Characterizing Historical Drought for the Upper Blue Nile Basin, Ethiopia. *Geosciences Journal*, MDPI, *8*(3), 81-106.

Baykasoglu, A. (2005). Data mining and an application in cement industry. *7th Academic Informatics Congress*, Gaziantep.

Cigizoglu, H.K. (2005). Generalized regression neural network in monthly flow forecasting. *Civil Engineering and Environmental Systems*, *22*(2), 71-84.

Caldas, C.H., Soibelman, L., & Han, J. (2002). Automated classification of construction project documents. *Journal of Computing in Civil Engineering*, *16*(4), 234-243.

Cavus, Y., & Aksoy, H. (2019). Spatial drought characterization for Seyhan River Basin in the Mediterranean Region of Turkey. *Water, MDPI*, *11*(7), 1331-1348.

Damle, C., & Yalcin, A. (2007). Flood prediction using time series data mining. *Journal of Hydrology*, *333*(2), 305-316.

Dawidowicz, J., Czapczuk, A., & Piekarski, J. (2018). The application of artificial neural networks in the assessment of pressure losses in water pipes in the design of water distribution systems. *Rocznik Ochorona Srodowiska*, *20*, 292-308.

Dawson, C.W., & Wilby, R. (1999). A comparison of artificial neural networks used for river flow forecasting. *Hydrology and Earth System Sciences*, *3*, 529-540.

Electrical Work Surveying Administration (EWSA) (2008). Water flows annual book. *General Directorate of Electrical Power Resources Survey and Development Administration, Hydraulic Study Management Department*, Ankara, Turkey, 433-470.

Fujihara Y., Simonovic P.S., Topaloglu, F., Tanaka, K., & Tsugihiro, W. (2008). An inverse modelling approach to assess the impacts of climate change in the Seyhan River Basin, Turkey. *Hydrological Sciences Journal*, *53*(6), 1121-1136.

Gumus, V., Soydan, N.G., Simsek, O., Akoz, M.S., & Kirkgoz, M.S. (2013). Comparison of different artificial neural networks for rainfall-runoff modeling. *Cukurova University Engineering and Architecture Journal*, *28*(1), 37-50.

Gumus, V., & Kavsut, M.E. (2013). Estimation of Missing Monthly Flow Data of Zamanti River-Ergenusagi Station. *Gazi University Journal Sci, Part C*, *1*(2), 81-91.

Gumus, V., & Algin, H.M. (2017). Meteorological and hydrological drought analysis of the Seyhan-Ceyhan River Basins, Turkey. *Meteorological Applications Journal*, *24*(1), 62-73.

Gumus, V. (2017). Hydrological drought analysis of Asi River Basin with streamflow drought index. *Gazi University Journal Sci, Part C*, *5*(1), 65-73.

Gumus, V., Yenigun, K., Toprak, Z.F., & Baci, N.O. (2018). Comparison of ANN, ANFIS and GEP methods in temperature-based evaporation estimation in Sanliurfa and Diyarbakir stations. *Dicle University Faculty of Engineering Journal*, *9*(1), 553-562.

Guo, S., Liao, X., Liu, F., & Zhu, Y. (2015). Collaborative Computing: Networking, Applications, and Worksharing, *11th International Conference*, CollaborateCom, Wuhan, China.

Hatami, P., Luo, L., Pei, L., Liu, X., Wilson, T., & Tan, P. N. (2018). *Predicting US Drought Monitor Drought Categories with multiple land surface models and machine learning.* In AGU Fall Meeting Abstracts.

Ikiel, C., & Ozyildirim, O. (2013). Rainfall forecasting using neural networks in Thrace. *2th International Balkan Annual Conference*, At Tirane.

Kamble V. B., & Deshmukh S. N. (2019). Comparison between Accuracy and MSE, RMSE by using Proposed Method with Imputation Technique. *Orient. J. Comp. Sci. and Technol*, *10*(4).

Kaur, A., & Sood, S. K. (2019). Cloud-Fog based framework for drought prediction and forecasting using Artificial Neural Network and Genetic Algorithm. *Journal of Experimental & Theoretical Artificial Intelligence*, 1-17.

Kaya, M., Keles, A.E., & Laptali Oral, E. (2013). Construction crew productivity prediction by using data mining methods. *Proceedings of the 4th World Conference on Learning, Teaching and Educational Leadership, Procedia-Social and Behavioral Sciences*, *141*, 1249-1253.

Kaya Keles, M. (2017). An overview: The impact of data mining applications on various sectors. *Technicki Glasnik*, *11*(3), 128-132.

Kaya Keles, M., & Keles, A.E. (2017). The place of data mining applications and heuristic optimization algorithms in construction management. *Cukurova University Engineering and Architecture Journal*, *32*(1), 235-242.

Keles, A.E., & Kaya, M. (2014). The analysis of the factors affecting the productivity in the wall construction of the using apriori data mining method. *Academic Informatics Congress, AIC Proceedings*, 831-836.

Keles, A.E. (2016). The overview of data mining application on construction sector and interpretation of economic impact. *Balkan Journal of Social Sciences, International Congress of Management Economy and Policy Special Issue*, 55-61.

Keskin, M.E., Terzi, O., Taylan, E.D., & Kucukyaman, D. (2007). *Scientific Research and Essays*, *6*(21), 4469-4477.

Keskin, M. E., Taylan, D., & Kucuksille, E. U. (2012). Data mining process for modeling hydrological time series. *Hydrology Research*, *44*(1), 78-88.

Kisi, O. (2006). Generalized regression neural networks for evapotranspiration modeling. *Hydrological Sciences Journal*, *51*(6), 1092-1105.

Konate, A.A., Pan, H., Khan, N., & Yang, J.H. (2015). Generalized regression and feed-forward back propagation neural networks in modelling porosity from geophysical well logs. *Journal of Petroleum Exploration and Production Technology*, *5*(2), 157-166.

Kubiak-Wójcicka, K., & Bak, B. (2018). Monitoring of meteorological and hydrological droughts in the Vistula basin (Poland). *Environmental Monitoring and Assessment*, *190*(11), 691.

Kumar, M.N., Murthy, C.S., Sesha, Sai., M.V.R., & Roy, P.S. (2009). On the use of standardized precipitation index (SPI) for drought intensity assessment. *Meteorological Applications*, *16*, 381-389. doi: 10.1002/met.136.

Kumar, P.S., Praveen, T.V., & Prasad, M. (2016). Artificial neural network model for rainfall-runoff -A case study. *International Journal of Hybrid Information Technology*, *9*(3), 263-272.

Kusiak, A., Wei, X., Verma, A.P., & Roz, E. (2013). Modeling and prediction of rainfall using radar reflectivity data: A data-mining approach. *IEEE Transactions on Geoscience and Remote Sensing*, *51*(4), 2337-2342.

Liaoa, C.W., & Perng, Y.H. (2008). Data mining for occupational injuries in the Taiwan Construction Industry. *Safety Science*, *46*(7), 1091-1102.

Lin, Y., Wen, H., & Liu S. (2018). Surface runoff response to climate change based on Artificial Neural Network (ANN) models: a case study with Zagunao catchment in Upper Minjiang River, Southwest China. *Journal of Water and Climate Change*, *10*(1), 158-166.

Loyeh, N.S., & Jamnani, M.R. (2017). Comparison of different rainfall-runoff models performance: A case study of Liqvan catchment, Iran. *European Water*, *57*, 315-322.

Machado, F., Mine, M., Kaviski, E., & Fill H. (2011). Monthly rainfall–runoff modelling using artificial neural networks. *Hydrological Sciences Journal*, *56*(3), 349-361.

McKee, T.B., Doesken, N.J., & Kleist, J. (1993). The relationship of drought frequency and duration of time scales. *8th Conference on Applied Climatology, American Meteorological Society*, Anaheim CA, 179-186.

Mehdiyev, N., Enke, D., Fettke, P., & Loos, P. (2016). Evaluating forecasting methods by considering different accuracy measures. *Procedia Computer Science*, *95*, 264-271.

Mishra, N., Soni, H. K., Sharma, S., & Upadhyay, A. K. (2018). Development and analysis of Artificial Neural Network models for rainfall prediction by using time-series data. *International Journal of Intelligent Systems and Applications*, *11*(1), 16.

Myronidis, D., Ioannou, K., Fotakis, D., & Dörflinger, G. (2018). Streamflow and hydrological drought trend analysis and forecasting in Cyprus. *Water Resources Management*, *32*, 1759-1776. doi: 10.1007/s11269-018-1902-z.

Nourani, V., Alami, M. T., & Aminfar, M. H. (2009). A combined neural-wavelet model for prediction of Ligvanchai watershed precipitation. *Engineering Applications of Artificial Intelligence*, *22*(3), 466-472.

Ozel, C., & Topsakal, A. (2014). Prediction of concrete compressive strength using data mining. *Cumhuriyet University Faculty of Science, Science Journal (CSJ)*, *35*(1), 43-57. ISSN: 1300-1949.

Patel, A.B., & Joshi, G.S. (2017). Modeling of Rainfall-Runoff Correlations Using Artificial Neural Network-A Case Study of Dharoi Watershed of a Sabarmati River Basin, India. *Civil Engineering Journal*, *3*(2), 78-87.

Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, *2*(1), 37-63.

Sattari, M.T., Yurekli, K., & Unlukara, A. (2011). Drought estimation by using artificial neural networks approach in Karaman province. *Journal of Agricultural Sciences*, *4*(1), 07-13.

Sattari, M.T., Mirabbasi, R., Sushab, R.S., & Abraham, J. (2018). Prediction of groundwater level in Ardebil Plain using support vector regression and M5 tree model. *Groundwater*, *56*(4), 515-679. doi: 10.1111/gwat.12620.

Sattari, M.T., & Sureh, F.S. (2019). Drought prediction based on standardized precipitation-evapotranspiration index by using M5 tree model. *International Civil Engineering and Architecture Conference (ICEARC)*, at Karadeniz Technical University.

Sezen, C., Bezak, N., Bai, Y., & Sraj, M. (2019). Hydrological modelling of karst catchment using lumped conceptual and data mining models. *Journal of Hydrology*, *576*, 98-110.

Shukla, S., & Wood, A.W. (2008). Use of a standardized runoff index for characterizing hydrologic drought. *Geophysical Research Letters*, *35*, L02405. doi: 10.1029/2007 GL032487.

Stachowski, P. (2010). Assessment of meteorological droughts on the postmining areas in the Konin Region. *Rocznik Ochorona Srodowiska*, *12*(1), 587-606.

Tabari, H., Nikbakht, J., & Hosseinzadehtalaei, P. (2013). Hydrological drought assessment in Northwestern Iran based on streamflow drought index (SDI). *Water Resources Management*, *27*(1), 137-151.

Tayyab, M., Zhou, J., Zeng, X., & Ikram, R.M.A. (2016). Discharge forecasting by applying artificial neural networks at the Jinsha River Basin, China. *European Scientific Journal*, *12*(9), 108-127. doi: 10.19044/esj.2016.v12n9p108.

Terzi, O (2012). Monthly rainfall estimation using data-mining process. *Applied Computational Intelligence and Soft Computing*, 20, 1-7. doi:10.1155/2012/ 698071.

The Ministry of Forestry & Water Affairs (MFWA) (2016). *The project of the sectoral water allocation plan of the Seyhan Basin*. General Directorate of Water Management, Ankara, Turkey.

Topcu E., & Seckin, N. (2016). Drought Analysis of the Seyhan Basin by Using Standardized Precipitation Index (SPI) and L-moments. *Journal of Agricultural Sciences*, *22*(2), 196-215.

Trafalis, T. B., Richman, M. B., White, A., & Santosa, B. (2002). Data mining techniques for improved WSR-88D rainfall estimation. *Computers and Industrial Engineering*, *43*(4), 775-786.

Tri, D.Q., Dat, T.T., & Truong D.D. (2019). Application of Meteorological and Hydrological Drought Indices to Establish Drought Classification Maps of the Ba River Basin in Vietnam. *Hydrology Journal, MDPI*, *6*(2), 49-68.

Tuncok, I.K. (2016). Drought planning and management: Experience in the Seyhan River Basin, Turkey. *Water Policy*, *18*(2), 177-209.

Turan, M.E., & Yurdusev, M.A. (2009). River flow estimation from upstream flow records by artificial intelligence methods. *Journal of Hydrology*, *369*(1), 71-77.

Turhan, E., Ozmen-Cagatay, H., & Cetin, A. (2016a). Modelling of rainfall-runoff relation with artificial neural network methods for Lower Seyhan Plain Sub-Basin and assessment in point of rainy-droughty terms. *Cukurova University Engineering and Architecture Journal, 31*(2), 227-241.

Turhan, E., & Ozmen-Cagatay, H. (2016b). Using of artificial neural network (ANN) for setting estimation model of missing flow data: Asi River-Demirköprü flow observation station (FOS). *Cukurova University Engineering and Architecture Journal*, *31*(1), 93-106.

Turhan, E., Tantekin, A., & Ozdil, N.F.T. (2016c). The evaluation of hydrological drought and energy efficiency relation in the context of pumped storage hydroelectric power plants (PSHPPs) issue: The case of Adana. *International Energy & Engineering Conference*, At Gaziantep.

Wilmot, C.G., & Cheng, G. (2003). Estimating future highway construction costs. *Journal Construction Engineering Management*, *129*(3), 272-279.

World Meteorological Organization (2012). *Standardized precipitation index user guide*. (WMO-No. 1090), Geneva, ISBN 978-92-63-11091-6.

Yurekli, K., Taghi Sattari, M., Anli, A. S., & Hinis, M. A. (2012). Seasonal and annual regional drought prediction by using data-mining approach. *Atmósfera*, *25*(1), 85-105.

Abstract

Proper water resources planning and management is based on reliable hydrological data. Missing rainfall and runoff observation data, in particular, can cause serious risks in the planning of hydraulics structures. Hydrological modeling process is quitely complex. Therefore, using alternative estimation techniques to forecast missing data is reasonable. In this study, two data-driven techniques such as Artificial Neural Networks (ANN) and Data Mining were investigated in terms of availability in hydrology works. Feed Forward Back Propagation (FFBPNN) and Generalized Regression Neural Networks (GRNN) methods were performed on rainfall-runoff modeling for ANN. Besides, Hydrological drought analysis were examined using data mining technique. The Seyhan Basin was preferred to carry out these techniques. It is thought that the application of different techniques in the same basin could make a great contribute to the present work. Consequently, it is seen that FFBPNN is the best model for ANN in terms of giving the highest R2 and lowest MSE values. Multilayer Perceptron (MLP) algorithm was used to predict the drought type according to limit values. This system has been applied to show the relationship between hydrological data and measure the prediction accuracy of the drought analysis. According to the obtained data mining results, MLP algorithm gives the best accuracy results as flow observation stations using SRI-3 month data.

Keywords:
artificial neural networks, drought analysis, data mining, Multilayer Perceptron, Seyhan Basin