S

Environmental Resources Research
Vol. 7, No. 1, 2019

GUASNR

Derivation of regression models for pan evaporation estimation

M. Jafari[1*], Y. Dinpashoh[2]
[1]PhD student, Department of Water Engineering, Faculty of Agriculture, University of Tabriz, Iran
[2]Associate Professor, Department of Water Engineering, Faculty of Agriculture, University of Tabriz, Iran

Received: September 2017 ; Accepted: September 2018

Abstract

Evaporation is an essential component of hydrological cycle. Several meteorological factors play role in the amount of pan evaporation. These factors are often related to each other. In this study, a multiple linear regression (MLR) in conjunction with Principal Component Analysis (PCA) was used for modeling of pan evaporation. After the standardization of the variables, independent components were obtained using the (PCA). The series of principal component scores were used as input in multiple linear regression models. This method was applied to four stations in East Azerbaijan Province in the North West of Iran. Mathematical models of pan evaporation were derived for each station. The results showed that the first three components in all four stations account for more than 90% of the data variance. Performance criteria, namely coefficient of determination (R^2) and root mean square error (RMSE), were calculated for models in each station. The results showed that in all the PCA-MLR models, the R^2 value was greater than 0.74 (significant at the 5% level) and the RMSE was less than 0.52 mm per day. In general, the results showed an improvement in the results using combination of PCA and MLR models for pan evaporation estimation.

Keywords: Climatic data, East Azerbaijan, Pan evaporation, Principal component analysis, Regression models, PCA-MLR

* Corresponding author; m.jafari.twone@gmail.com

Introduction

Evaporation is one of the main components of the hydrological cycle and water balance in natural and agricultural ecosystems. One of the statistical methods used in evaporation models is multiple linear regressions that finds probable relationship between dependent variable and independent variables. Several factors are involved in empirical evaporation models the most important of which are air temperature, wind speed, relative humidity, solar radiation and altitude of site (Chow *et al.* 1988). Studies conducted by researchers such as Bruton *et al.* (2000), Shirsath and Singh (2010), Shirgure (2011), Googhari (2012) and Shirgure and Rajput (2012) used the ANN and MLR methods with acceptable results. Shirsath and Kumar (2009), predicted daily pan evaporation using artificial neural network (ANN), MLR, Penman, Stephen and Stuart models. Almedeij (2012) predicted daily and monthly evaporation by MLR in Kuwait. For this purpose, parameters including air temperature, relative humidity and wind speed for 17 years (from January 1993 to December 2009) were used in a desert area. Finally, a linear relationship was created between evaporation, air temperature and relative humidity. The results showed that the model created with these parameters had high correlation with the observed data. Ladlani *et al.* (2013) modeled the daily evapotranspiration in the Mediterranean region of Algeria using neuro-fuzzy and MLR methods. The data used included mean daily relative humidity, number of sunshine hours, maximum, minimum and average air temperature and wind speed. The results showed that the performance of both models to predict evapotranspiration was acceptable. Malik *et al.* (2013) and Kishi (2009), used ANN and MLR methods to estimate the pan evaporation. The results of these two studies represent the good performance of MLR model. Malik and Kumar (2015) simulated daily pan evaporation by ANN, MLR and neuro-fuzzy methods in the area of Pantnagar (India). The comparison of models by R^2 and RMSE indicated ordered performances from ANN to neuro-fuzzy and the MLR, methods respectively. Eskafi Noghani *et al.* (2008) estimated pan evaporation by MLR method using meteorological parameters in Gorganrood basin (Iran). They compared results with corresponding measured values from evaporation pan. The results showed that output of MLR method in this area was almost equal to the measured values and therefore, pan evaporation was estimated accurately. Another multivariate statistical method widely used today is the combination of principal component analysis with multiple linear regression (MLR-PCA) (Tianxiao *et al.* 2009). PCA method can establish a linear relationship between the set containing the large number of variables and a limited number of principal components. The main purpose of PCA is to find a few principal components that justify a large percentage of the observed data variances. In this method, the variable vectors (with n elements) are converted into other vectors containing a smaller number of independent principal components. Kovoor and Nandagiri (2007) used the MLR and PCA methods to predict the daily pan evaporation based on data from four stations with different climates in India. They used 3 to 6 variables in MLR and each of the 6 variables used as inputs in the method of PCA. The results showed that the predicted evaporation with this combined method was very close to observations. Tianxiao *et al.* (2009) studied the impact of components obtained from PCA method on prediction of pan evaporation in the Arctic. They used data including air temperature, relative humidity, wind speed, actual vapor pressure and solar radiation. The results showed that the change in evaporation is the result of the combined effect of numerous metrological factors that interact with each other. Other researchers have also applied PCA in hydrological studies. In most of these research, the ability of PCA in modeling of hydrological processes has been proven (Yu and Yung 1996; Baeriswyl and Rebetez 1997; Bonaccors *et al.*, 2003; Neal and Phillips 2009; Raziei and Azizi 2009; Ghorbani Aghdam *et al.*, 2012; Fallahi *et al.*, 2012; Ghorbani Aghdam *et al.*, 2013). Konishi and Rao

(2014), conducted the PCA method on meteorological data (such as maximum, minimum and mean air temperatures and dew point). They obtained similar statistical distributions by reducing the number of data dimensions. Nazemosadat and Shirvani (2005) studied the Persian Gulf Sea surface temperature changes using the PCA and MLR. The first four components described 73.5% of the total variance and were introduced as principal components. The results showed that winter temperature of the Persian Gulf Sea surface is particularly related to the temperature of water in the last winter. Seifi *et al.* (2011) studied the reference evapotranspiration in Kerman station using the hybrid method MLR–PCA and examined the relative importance of the effected variables using factor analysis. The results of their study showed that radiation, relative humidity, sunshine hours, and maximum and minimum air temperature are the more important factors for evapotranspiration in this region. They reduced the number of variables using the PCA in Kerman station. Shirvani and Nazemosadat (2012) divided Iran into homogeneous precipitation regions using PCA. This was implemented for grouping of stations in terms of monthly precipitation,. Using this method, about 96% of the variance in the observed time series was accounted by few components. They divided the whole country into six homogeneous precipitation regions. Asakereh and Bayat (2013) conducted PCA for the regionalization of annual rainfall of Zanjan Province (Iran). The results showed that the first four principal components described about 95% of the annual rainfall changes. Sheikholeslami *et al.* (2014) developed the evapotranspiration in Mashhad station using hybrid method PCA-MLR. They used data from Mashhad station (semi-arid) on a daily scale in the period 1991- 2005 and examined the effects of parameters including air temperature (maximum, minimum and mean), relative humidity, sunshine hours and wind speed at 2 m height on evaporation. As a result, air temperature (minimum, mean and maximum) and relative humidity for evapotranspiration were found to be more important than other variables (wind speed and sunshine hours). Finally, values of R^2 was obtained for MLR-PCA and MLR methods with 0.903 and 0.89, respectively. It seems that, in Iran, no comprehensive study has been conducted so far in the field of pan evaporation data modeling using the combination of multiple linear regression and principal components analysis (MLR-PCA). The objectives of this study include i) using principal component analysis to reduce meteorological data dimensions, ii) modeling daily pan evaporation using MLR- PCA and iii) modeling daily pan evaporation using MLR.

Methodology

East Azerbaijan Province located in the north west of Iran was selected as the study area. The area of East Azerbaijan is approximately 45,491 square km and its geographical position is 45° 7′ to 48° 20′ eastern longitudes and 36° 45′ to 39° and 26′ northern latitudes. The average amount of precipitation in the region is 250 to 400 mm/year and the average of pan evaporation is approximately 1700 mm/year. Winter and spring are the rainy seasons and the highest intensity of rainfall occurs in the spring. The fall season is in the third rank of rainfall. Figure 1 shows the geographical location of East Azerbaijan Province and the selected stations. Table 1 lists the details of the selected stations.

Table 1. Details of the climatic stations selected in this study

Station	Latitude (N)	Longitude (E)	Altitude (M)	Establishment year	Data period
Tabriz	38 05	46 17	1364	1951	1992- 2012
Jolfa	38 56	45 36	736	1985	1992- 2012
Miane	37 27	47 42	1110	1978	1992- 2012
Maragheh	37 01	46 10	1344	1983	1992- 2012

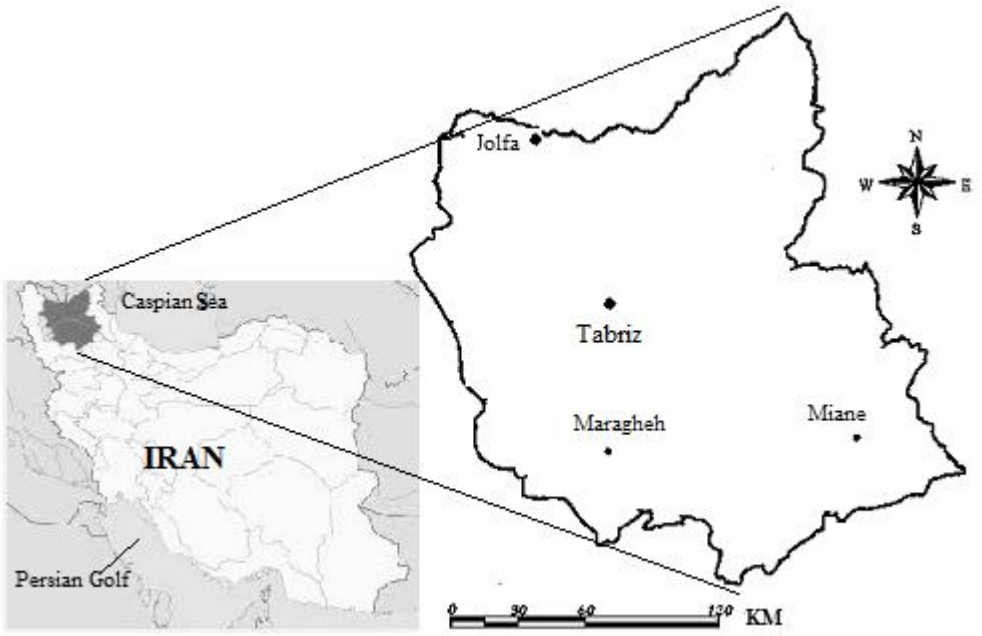


Figure 1. Location of the East Azerbaijan Province and the selected stations in the study

Data pre- processing

In order to meet the objectives of this study, the MLR and MLR- PCA techniques were used separately to develop regression models relating daily pan evaporation (E_{pan}) to various meteorological variables. Separate regression models were fitted using the historical records of climatic variables at the four selected synoptic stations (Tabriz, Jolfa, Miane and Maragheh). In all cases two regression models applied to climate datasets of four mentioned stations. Daily pan evaporation (E_{pan}) values were considered to be the response (dependent) variable and corresponding daily average of maximum air temperature (T_{max}), minimum air temperature (T_{min}), maximum relative humidity (RH_{max}), minimum relative humidity (RH_{min}), number of sunshine hours (n), wind speed (w) and dew point temperature (T_{dew}), were considered to be the regressors (independent). In addition, all regression modeles were developed using 75% of available daily records (in calibration phase) and subsequently tested using the remaining 25% (in validation phase).

Multiple Linear Regression (MLR)

The general form of the MLR model is given by

$$Y = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_p x_p + e \quad (1)$$

In which Y: response or criterion variable; x_i (i=1,2,…,p): predictor variables; p: number of predictor variables, and a_i (i=1,2,…,p) are the regression coefficients (maidment 1993).

The functional relationship between dependent and independent variables can be stated with matrix notation as follows:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (2)$$

where **Y** is an output vector of size n×1; **X** is an input matrix of size n ×(p + 1); **β** is a coefficient vector of size (p + 1) ×1 and **ε** is an error vector of size n×1. Eq. (2), can be written in the extended form as:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & . & x_{1p} \\ 1 & x_{21} & x_{22} & . & . & . & x_{2p} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & . & . & . & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix} \quad (3)$$

The regression parameter coefficients vector **β** can be estimated as below;

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}) \quad (4)$$

where **X'** is the transpose of **X**. To calculate the inverse of (**X´ X**), it is necessary that determinant of eq (4), is not be zero (Bowker and Lieberman 1972). Multiple linear regression equations of the form specified by Eq. (1) were developed using the forward best regression module available in Excel STAT software.

Principle Component Analysis (PCA)

In order to examine the relationships among a set of *p* correlated variables, it may be useful to transform the original set of variables to another new set of uncorrelated variables called principal components. These new variables are linear combinations of the original variables and are derived in decreasing order of importance so that, for example, the first principal component accounts for the largest variance of the original data. PCA originated in some work by Karl Pearson around the turn of the century, and was further developed in the 1930s by Harold Hotelling (Chatfield and Collins, 1980). The usual objective of the analysis is to see if the first few components account for most of the variation in the original data. In other words, if some of the original variables are highly correlated, they are effectively 'having the same information' and there may be near-linear constraints on the variables. This method will simply find components which are close to the original variables but arranged in decreasing order of variance (liu *et al.* 2003). As a result, the information of original variables was exhibited by derived principal components and don't waste aspects of data's information (Konishi and Rao 2014). The PCA can be explained as four below stages:

A) Calculation of KMO[1] factor

KMO value varies between zero and one. This factor was calculated using the simple and partial correlation coefficient according to the equation (5).

$$KMO = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^{\,2}}{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^{\,2} + \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}^{\,2}} \quad , \quad \forall (i \neq j) \quad (5)$$

In this equation, r_{ij} is simple correlation coefficient and a_{ij} is partial correlation coefficient between variables i and j. In the cases of the value of this factor was greater than 0.5, there is the possibility of implementing this approach (Shrestha and Kazama, 2007).

B) Standardization of input variables

Standardization of input data performed based on the following equation so that they have zero mean and standard deviation of one (Salas 1993).

$$Z = \frac{x - \mu}{\delta} \quad (6)$$

In this equation, Z standardized amounts of data, x observed data, μ and σ are the mean and standard deviation of the data.

C) Calculation of correlation matrix (R)

The correlation matrix (**R**), is a symmetric matrix that shows the pairwise correlation between P input variables. The diagonal elements of this matrix is equal to **1** and the others denoted by r_{ij} (correlation coefficient between input variables i and j).

D) Calculation of Eigenvalues (λ_i) and Eigenvectors

Suppose I_p is an identity matrix with the same dimensions with the dimensions of the matrix **R** (the p × p). By solving the following equation, eigenvalues i.e. $\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_p]^T$, can be obtained as follows (Anton 2005):

$$\left|\mathbf{R} - \lambda_i . \mathbf{I}_p\right| = \mathbf{0} \quad , \quad \forall (i = 1, 2, \ldots, p) \quad (7)$$

The symbol |. | in the eq. 7 denotes the determinant of the matrix. Eigenvalues were obtained in descending order $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, so that the sum of the eigenvalues should be equal to the number of variables p. The first principal component, PC_1 is found by choosing λ_1 so that λ_1 has the largest possible variance. The second principal component is found by choosing λ_2 so that PC_2 after PC_1

1- Kaiser- Meyer- Olkin

accounts the largest variance and is uncorrelated with PC_1. Similarly, other components (PC_3 ,…. , PC_p) were derived so that they are uncorrelated with eachothers and the variance of PC_3 is greater than PC_4 and so on. Scree plot is the main tool for determining the number of principal components. In this method the boundary between principal and redundant components is where the scree plot curve tend to be a horizontal line. In such a point the eigenvalues do not change considerably with increasing the number of components. The following equation used to calculate the eigenvectors corresponding to λ_i (Anton 2005):

$$(\lambda_i \mathbf{I} - \mathbf{R})\mathbf{X}^* = \mathbf{0} \quad , \quad I = 1,2,..,P \tag{8}$$

where $\mathbf{X}^* = [x_1^* ... x_p^*]^T$ is the eigenvector corresponding to eigenvalues λ_i. In the above equation λ_i, is the known value whereas elements of $\mathbf{X}^*$ are unknown.

Performance criteria

In this study, the performance of the models (MLR-PCA and MLR) evaluated using the following criteria: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Coefficient of Determination (R^2), and Durbin- Watson (DW) statistics. The two models are compared on the basis of statistical error criteria. They are defined by the following equations (Salas 1993):

$$RMSE = \left(\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2\right)^{0.5} \tag{9}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|p_i - o_i| \tag{10}$$

$$R^2 = \frac{\left(\sum_{i=1}^{N}(P_i - \bar{P})(O_i - \bar{O})\right)^2}{\sum_{i=1}^{N}(P_i - \bar{P})^2 \sum_{i=1}^{N}(O_i - \bar{O})^2} \tag{11}$$

In these equations, N: number of data, P_i: i_{th} observed value, O_i: i_{th} calculated value, $\bar{P}$ and $\bar{O}$ are the average of the observed and calculated values of pan evaporation, repectively. For ideal data modeling, RMSE (mm/day) and MAE (mm/day), should be closer to zero, but value of R^2 should be approach to 1 as closely as possible. The last criterion used in this study is the Durbin- Watson statistic (DW) which calculated as follows (Bowker and Lieberman 1972):

$$D = \frac{\sum_{t=2}^{N}(e_t - e_{t-1})^2}{\sum_{t=1}^{N} e_t^2} \tag{12}$$

where e_t, is the t-th residual value of model. The critical values for Durbin-Watson is significant for more than 100 data's (at the significant level of 5 %), are between 1.44 to 1.78.

Results and Discussion

KMO was used to determine the feasibility of principal component analysis. The values of KMO for stations Tabriz, Jolfa, Maragheh, and Miane, are respectively, 0.699, 0.618, 0.668 and 0.622 were confirmed PCA method. After standardization of the values of input variables, correlation matrix (R) with dimensions of 7 × 7 was formed. Table 2 shows the elements of the mentioned matrix for the selected stations. Using the Equation 8, seven eigenvalues and their corresponding eigenvectors was obtained.

Table 2. Correlation coefficients matrix for the selected stations

		Variables							
Station		T_{min}	T_{max}	RH_{max}	RH_{min}	W	T_{dew}	n	E_{pan}
Tabriz	T_{min}	1.000	0.930	-0.630	-0.529	0.360	0.602	0.552	0.788
	T_{max}		1.000	-0.767	-0.707	0.227	0.488	0.622	0.768
	RH_{max}			1.000	0.912	-0.144	0.138	-0.587	-0.591
	RH_{min}				1.000	-0.092	0.145	-0.541	-0.503
	w					1.000	0.234	0.155	0.388
	T_{dew}						1.000	0.208	0.410
	n							1.000	0.589
	E_{pan}								1.000
Jolfa	T_{min}	1.000	0.857	-0.696	-0.381	0.605	0.751	0.449	0.844
	T_{max}		1.000	-0.647	-0.675	0.335	0.634	0.629	0.744
	RH_{max}			1.000	0.552	-0.562	-0.151	-0.456	-0.755
	RH_{min}				1.000	-0.101	0.044	-0.618	-0.435
	w					1.000	0.310	0.260	0.651
	T_{dew}						1.000	0.243	0.488
	n							1.000	0.501
	E_{pan}								1.000
Miane	T_{min}	1.000	0.873	-0.550	-0.416	0.453	0.621	0.396	0.784
	T_{max}		1.000	-0.708	-0.719	0.269	0.382	0.631	0.767
	RH_{max}			1.000	0.741	-0.191	0.198	-0.487	-0.616
	RH_{min}				1.000	-0.095	0.245	-0.656	-0.498
	w					1.000	0.309	0.159	0.486
	T_{dew}						1.000	0.079	0.360
	n							1.000	0.524
	E_{pan}								1.000
Maragheh	T_{min}	1.000	0.929	-0.659	-0.496	0.379	0.507	0.456	0.794
	T_{max}		1.000	-0.756	-0.676	0.264	0.372	0.624	0.790
	RH_{max}			1.000	0.772	-0.122	0.179	-0.543	-0.651
	RH_{min}				1.000	0.030	0.309	-0.634	-0.530
	w					1.000	0.372	0.099	0.350
	T_{dew}						1.000	0.059	0.277
	n							1.000	0.524
	E_{pan}								1.000

Table 3 shows the variances explained by components extracted from PCA. Table 4 shows the values of eigenvectors (Loading), which gives coefficients of meteorological parameters. As shown in Table 3, the first component, for example for Tabriz station, which is 3.1788, justified 55.53% of the total variance in the data series. The second and third eigenvalues explained 22.44% and 12.16% of total variances. These three components justified overall, about 90% of total variance of observations. Therefore, the first three components considered as the principal components, in this study. According to the scree plot (Figure 2), the slope of scree plot decreased fast as the numbers of components changed from 3 to 4. Noori Gheidari (2010), in Lake Urmia, has chosen the three first components that justified 87.5% of the data's variance by principal components. To calculate the first principal component scores eigenvectors should be multiplied by the standardized variables of the meteorological parameters.

Table 3. Total variance explained by each of the components extracted from PCA

Station	Component	Initial eigenvalues		
		Eigenvalue	% of variance	Cumulative %
Tabriz	1	3.887	55.530	55.530
	2	1.557	22.247	77.777
	3	0.851	12.159	89.937
	4	0.510	7.287	97.225
	5	0.132	1.888	99.113
	6	0.040	0.574	99.688
	7	0.021	0.311	100
Jolfa	1	3.945	56.359	56.359
	2	1.347	19.248	75.607
	3	0.917	13.104	88.712
	4	0.484	6.915	95.627
	5	0.230	3.337	98.964
	6	0.042	0.600	99.565
	7	0.030	0.434	100
Miane	1	3.650	52.150	52.150
	2	1.715	24.510	76.660
	3	0.772	11.033	87.694
	4	0.567	8.105	95.799
	5	0.209	2.988	98.788
	6	0.05	0.734	99.522
	7	0.033	0.477	100
Maragheh	1	3.720	53.155	53.155
	2	1.751	25.021	78.177
	3	0.709	10.140	88.317
	4	0.539	7.707	96.025
	5	0.188	2.691	98.716
	6	0.059	0.843	99.560
	7	0.030	0.439	100

For example, in the case of Tabriz, the PC_1 obtained by multiplication of the standardized maximum air temperature, by 0.486, the standardized minimum air temperature, by 0.455, and in the same way, the standardized values of other variables by their coefficients and summing the resultant values. This is done similarly for the PC_2, PC_3 and so on.

For example, in the case of Tabriz station, the first three principal components (selected components), calculated using the following formula:

$$PC_1 = (0.486 \times T_{\max}) + (0.455 \times T_{\min}) + (-0.434 \times RH_{\max}) + (-0.407 \times RH_{\min}) + (0.378 \times n) + (0.164 \times W) + (0.176 \times T_{dew}) \quad (13)$$

$$PC2 = (0.102 \times T_{\max}) + (0.272 \times T_{\min}) + (0.375 \times RH_{\max}) + (0.411 \times RH_{\min}) + (-0.057 \times n) + (0.331 \times W) + (0.702 \times T_{dew}) \quad (14)$$

$$PC_3 = (-0.166 \times T_{\max}) + (-0.072 \times T_{\min}) + (-0.086 \times RH_{\max}) + (-0.057 \times RH_{\min}) + (-0.115 \times n) + (0.92 \times W) + (-0.311 \times T_{dew}) \quad (15)$$

The number of members for each set of PCs calculated by equations 13 to 15, is equal to the number of observation days. These three series have no correlation with each other, and so, were used as input data for the multiple linear regressions.

Derivation of (MLR- PCA) models

At each station, obtained series of PC_1, PC_2, PC_3 (Equations 13 to 15), were used as input data in multiple linear regression model. For example, in calibration phase (using the first 80% of data), the multiple linear regression model for estimation of pan evaporation in Tabriz station obtained as follows:

$$E = 0.022 + 0.401(PC_1) + 0.159(PC_2) + 0.067(PC_3) \quad (16)$$

in which PC_1, PC_2 and PC_3 calculated from (13) to (15), respectively. Similarly for the test data (remaining 20% of data), the multiple linear regression model were presented as follows:

$$E = -0.067 + 0.394(PC_1) + 0.125(PC_2) + 0.151(PC_3) \quad (17)$$

The MLR- PCA models were derived for all the stations in a some way. Find results summarized in Table 5. In order to compare the results of MLR-PCR method with that of the MLR, all values of inpu variables

were entered as input for MLR model. The results were presented in Table 6.

Table 4. Coefficient of variables (eigenvectors or loadings) derived using the principal component analysis

		Eigenvectors						
Station		PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7
Tabriz	T min	0.455	0.271	-0.072	-0.222	0.478	0.648	-0.119
	Tmax	0.486	0.101	-0.166	-0.213	0.042	-0.372	0.734
	Rh max	-0.434	0.375	-0.086	0.161	-0.294	0.510	0.538
	Rh min	-0.407	0.411	-0.057	0.186	0.711	-0.346	0.036
	w	0.164	0.331	0.919	0.072	-0.075	-0.062	0.046
	T dew	0.176	0.702	-0.311	-0.038	-0.413	-0.233	-0.390
	n	0.378	-0.057	-0.115	0.915	0.022	0.046	0.019
Jolfa	T min	0.463	0.280	-0.011	-0.196	-0.039	0.738	-0.346
	T max	0.469	-0.031	-0.294	-0.216	0.102	-0.590	-0.534
	Rh max	-0.400	0.159	-0.450	0.347	0.605	0.160	-0.313
	Rh min	-0.322	0.596	0.103	0.207	-0.560	-0.173	-0.376
	w	0.298	0.292	0.675	0.366	0.454	-0.167	-0.003
	Tdew	0.295	0.582	-0.449	0.044	0.022	-0.122	0.594
	n	0.351	-0.337	-0.203	0.783	-0.314	0.082	-0.027
Miane	T min	0.443	0.338	-0.079	-0.304	0.027	0.700	-0.313
	T max	0.501	0.062	-0.213	-0.143	-0.174	-0.068	0.802
	Rh max	-0.415	0.302	-0.188	0.4290	-0.585	0.380	0.168
	Rh min	-0.410	0.399	-0.012	-0.047	0.698	0.188	0.382
	w	0.211	0.361	0.865	0.238	-0.083	-0.063	0.089
	T dew	0.143	0.685	-0.355	0.078	-0.010	-0.550	-0.272
	n	0.381	-0.173	-0.191	0.798	0.361	0.134	-0.042
Maragheh	T min	0.457	0.268	-0.122	-0.315	-0.014	0.668	-0.396
	T max	0.498	0.111	-0.173	-0.146	-0.102	-0.002	0.823
	Rh max	-0.441	0.247	-0.183	0.329	-0.631	0.415	0.175
	Rh min	-0.407	0.394	-0.073	0.001	0.721	0.282	0.267
	w	0.156	0.467	0.844	0.182	-0.085	-0.031	0.041
	t dew	0.107	0.678	-0.426	0.080	-0.058	-0.529	-0.239
	n	0.382	-0.141	-0.144	0.854	0.242	0.142	-0.060

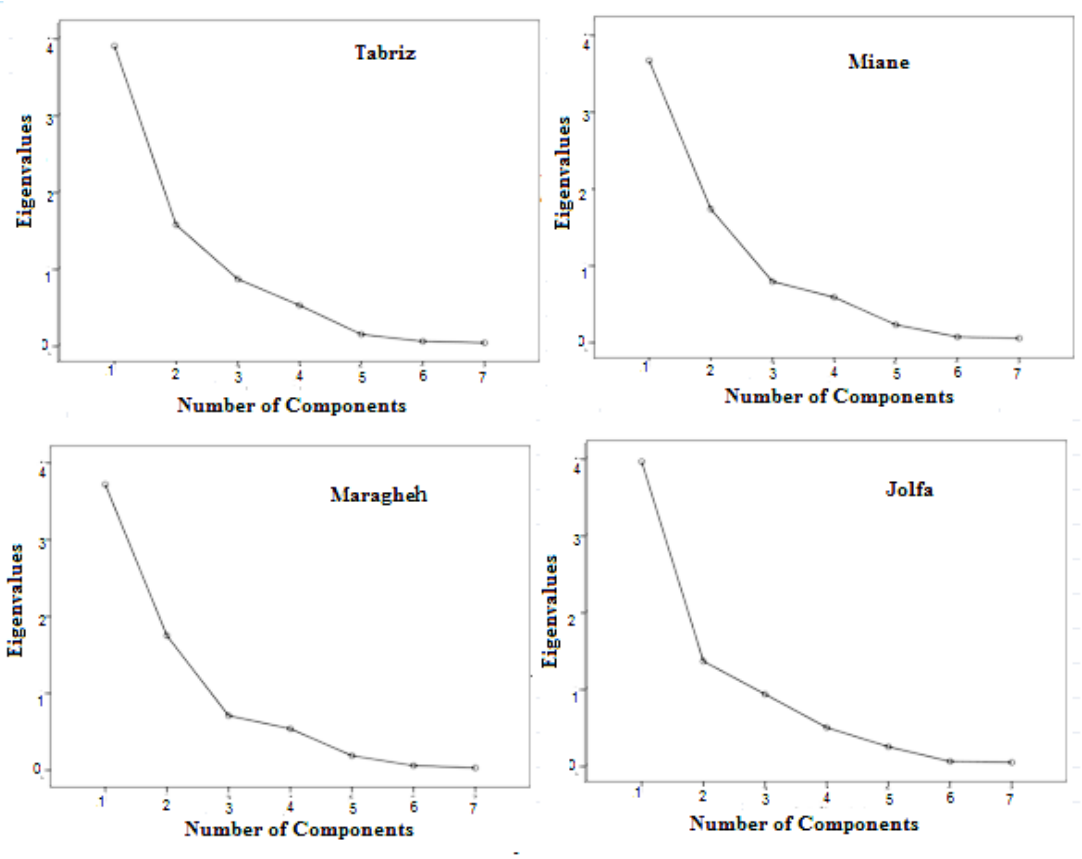


Figure 2. Scree plots obtained using the PCA for the selected stations

Table 5. Models derived using the principal components regression

Station	No. of Component	Training phase		Testing phase	
		Variables	Regression coefficients	Variables	Regression coefficients
Tabriz	3	Constant	0.022	Constant	-0.067
		PC1	0.401	PC1	0.394
		PC2	0.159	PC2	0.125
		PC3	0.067	PC3	0.151
Jolfa	3	Constant	0.039	Constant	-0.101
		PC1	0.435	PC1	0.420
		PC2	0.105	PC2	0.088
		PC3	0.202	PC3	0.219
Miane	3	Constant	0.022	Constant	-0.222
		PC1	0.414	PC1	0.424
		PC2	0.118	PC2	0.166
		PC3	0.037	PC3	0.098
Maragheh	3	Constant	0.011	Constant	0.036
		PC1	0.423	PC1	0.415
		PC2	0.151	PC2	0.168
		PC3	0.113	PC3	0.159

Table 6. Final models derived using the suitable multiple linear regression for selected stations

Station	Variables	Regression coefficients	R^2	D
Tabriz	Constant	-0.731	0.65	1.48
	T_{min}	0.481		
	n	0.299		
Jolfa	Constant	9.22	0.76	1.46
	T_{min}	0.525		
	RH_{max}	-0.107		
Miane	Constant	-4.288	0.67	1.58
	T_{max}	0.395		
	w	0.722		
Maragheh	Constant	-0.548	0.66	1.47
	T_{min}	0.504		
	n	0.258		

Table 7 represents coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE) and the Durbin-Watson statistic (D) for models during validation. RMSE values for MLR-PCA never exceeded at any station from 0.52 mm/day. R^2 values of MLR-PCA are also varied from 0.74 to 0.82 at the Miane and Jolfa stations, respectively. MAE values for MLR-PCA are also varied from 0.64 to 0.85 mm/day at the Maragheh and Miane stations, respectively. The Durbin-Watson statistic, which is less than 2 shows the validity of the models. It can be concluded that the MLR-PCA models for all four stations were superior to MLR models.

Table 7. Performance of regression models during validation

Station	MLR- PCA				MLR			
	R^2	RMSE (mm/day)	MAE (mm/day)	D	R^2	RMSE (mm/day)	MAE (mm/day)	D
Tabriz	0.76	0.45	0.70	1.62	0.73	2.20	1.70	1.50
Jolfa	0.82	0.40	0.65	1.75	0.79	2.41	1.85	1.80

Miane	0.74	0.52	0.85	1.65	0.70	2.28	1.70	1.65
Maragheh	0.75	0.50	0.64	1.60	0.71	2.22	1.67	1.57

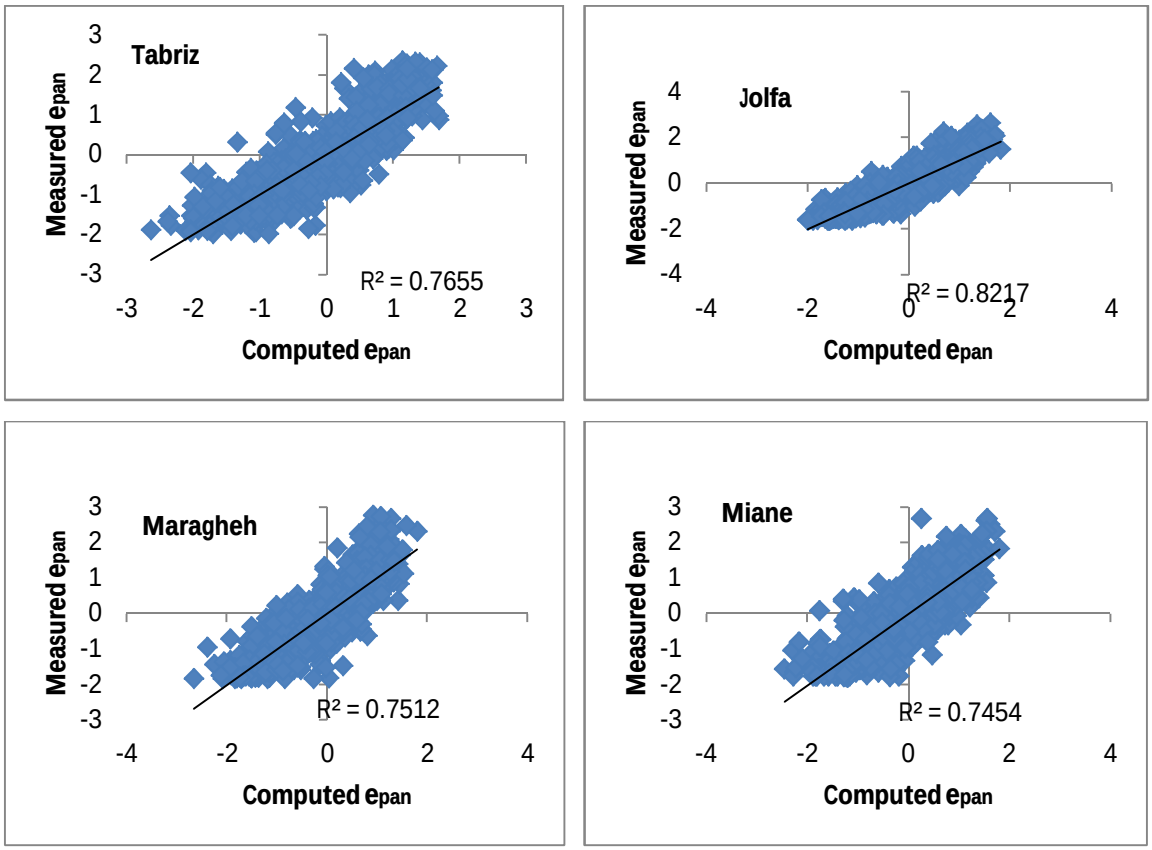
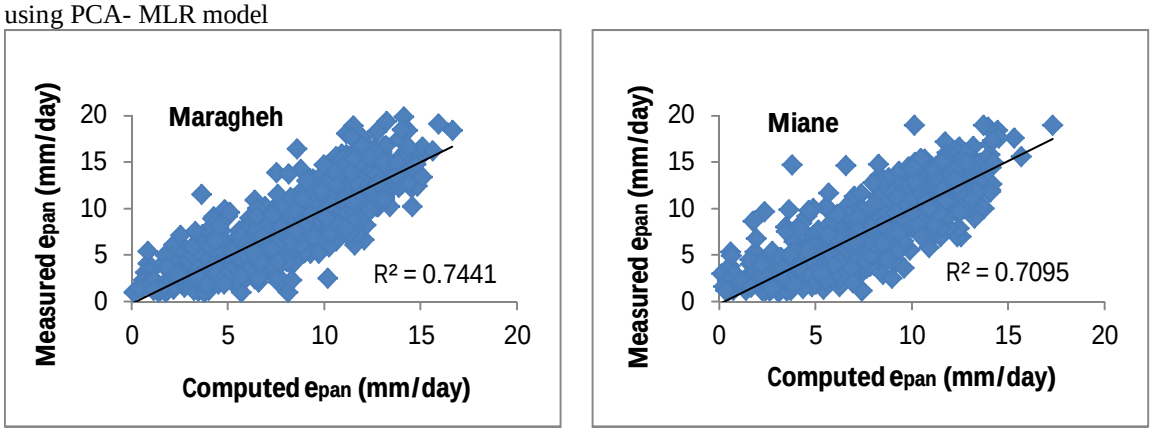


Figure 3. Comparison of the observed daily pan evaporation (Standardized data) with those computed using PCA- MLR model



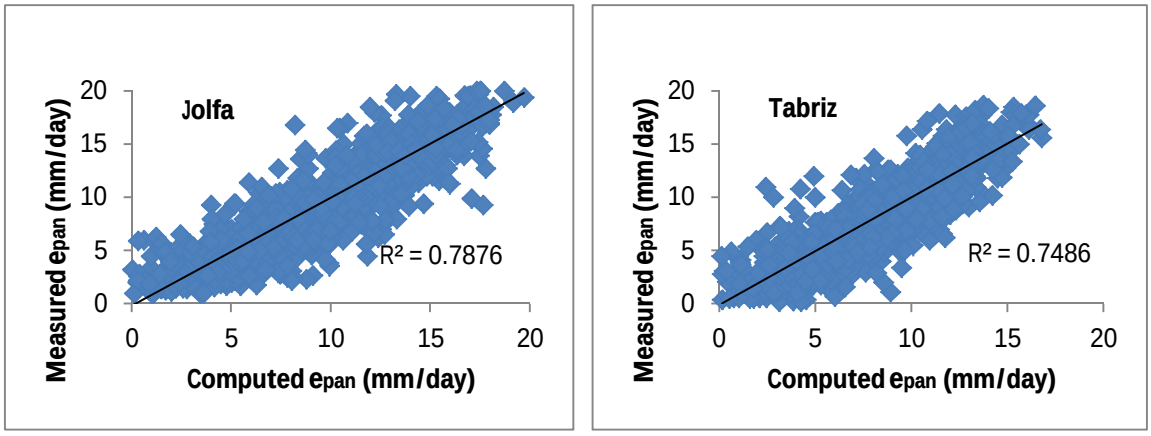


Figure 4. Comparison of the observed daily pan evaporation with those computed using MLR model

Figures 3 and 4 show the scatter plot of the observed and computed values of daily pan evaporation (Standardized data) using the PCA-MLR and MLR models, respectively. As it can be inferred from these figures, there are scatter points located around the line 1:1 for both models. However, the cloud of points in the model MLR- PCR, is slightly thinner than MLR. Also in Figure 4 on all stations there are few points near the horizontal axis implying a very high valued calculated for pan evaporation against the observed one. It seems that these points are related to measurement errors. Seifi *et al.* (2011) studied the reference evapotranspiration of Kerman station using a hybrid model (MLR-PCA), and acquired R^2= 0.8 and RMSE= 0.1 using the two first principal components (explaining 80% of the total variance). Tianxiao *et al.* (2009) chose three principal components, which explained 97% of the total variance.

Conclusion

Evaporation is considered as one of the limiting factors in utilization of water resources. In this study, combination of MLR-PCR method, MLR method and the relative importance of the effective variables in pan evaporation were evaluated. For this purpose, we used data of four stations namely Tabriz, Jolfa, Maragheh, and Miane in East Azerbaijan Province (IRAN), and seven meteorological parameters were considered in the analysis. These included the maximum and minimum air temperature, dew point temperature, maximum and minimum relative humidity, number of sunshine hours and wind speed. Considering certain inherent limitations in application of datasets with significant degree of correlation between the predictor variables (multicollinearity), the principal component analysis (PCA) approach has found wide application in development of empirical models for estimation of evaporation/ evapotranspiration rates from climatic observations and also in calculating several other climate- dependent parameters. In previous studies, such as Kovoor and Nandagiri (2007), Seifi *et al.* (2011), and Sheikholeslami *et al.* (2014) principal component analysis was used to estimate evapotranspiration and to overcome the effects of correlation between the input variables (multicollinearity). These also used the performance statistics such as R^2 and RMSE. To avoid the effects of multicollinearity, we used principal component analysis to predict pan evaporation in East Azerbaijan Province. The PCA method offers the components with controlled coefficients and the correlation between the components created in this method is exactly zero and there is no multicollinearity. Results showed acceptable performance to estimate the pan evaporation in these stations for MLR-PCA model compared with MLR model. While accepting that our conclusions are specific to our datasets, the finding of this study highlight the need to more tests on the advantages offered by the MLR-PCA regression approach, relative to the popular MLR approach.

References

Almedeij, J. 2012. Modeling Pan Evaporation for Kuwait by Multiple Linear Regression. Journal of the Scientific World. 9, 10-11.

Anton, H. 2005. Elementary Linear Algebra: (11th Edition): John Wiley& sons.

Asakereh, H., and Bayat, A. 2013. The Analysis of the Trend and the Cycles of Annual Precipitation Characteristics of Zanjan. Journal of Geography and Planning. 17(45), 121- 142. (In Persian)

Baeriswyl, P.A., and Rebetez, M. 1997. Regionalization of precipitation in Switzerland by means of principal component analysis. Journal of Theoretical and Applied Climatology. 58, 31-41.

Bonaccors, B., Bordi, I., Cancellier, A., Rossi, G., and Sutera, A. 2003. Spatial variability of drought: An analysis of the SPI in Sicily. Journal of Water Resources Management. 17, 273–296.

Bowker, H., and Lieberman, G.J. 1972. Engineering Statistics: Prentice-Hall: 852P.

Bruton, J.M., Mcclendon, R.W., and Hoogenboom, G. 2000. Estimating daily pan evaporation with artificial neural networks. Trans ASAE. 43(2), 491–496.

Chatfield, C., and Collins, A.J. 1980. Introduction to Multivariate Analysis: Chapman & Hall: London and New York.

Chow, V.T., Maidment, D.R., and Mays, L.W. 1988. Applied Hydrology. New York: McGraw Hill.

Eskafi Noghani, M., Meftah Halghi, M., and Mosaedi, A. 2008. Offering a regression model for evaporation losses using the measured meteorological parameters. Proceeding of 3rd Iran Water Resources Management Conference in Tabriz University, Tabriz, Iran. October 14-16. (In Persian)

Fallahi, B., Fakheri Fard, A., Dinpajooh, Y., and Darbandi, S. 2012. Regionalization of Northwest Iran Based on Daily Rainfalls and Rain's Time Intervals Using PCA, Ward and K-mean Methods. Journal of Water and Soil. 26(4), 979- 989. (In Persian)

Ghorbani Aghdam, M., Dinpashoh, Y., Fakheri Fard, A., and Darbandi, S. 2012. Regionalization of Urmia Lake Basin from the View of Drought Using Factor Analysis. Journal of Water and Soil. 26(5), 1268- 1276. (In Persian)

Ghorbani Aghdam, M., Dinpashoh, Y., and Mostafaeipour A. 2013. Application of factor analysis in defining drought prone areas in Lake Urmia Basin. Journal of Natural Hazards. 69, 267–277.

Googhari, S.K. 2012. Daily pan evaporation estimation using a neuro-fuzzy-based model. Journal of Agricultural Science and Technology. 2, 223–228.

Kishi, O. 2009. Modeling monthly evaporation using two different neural computing techniques. Irrigation Science. 27, 417–430.

Konishi, S., and Rao, C.R. 2014. Principal component analysis for multivariate familial data. Journal of Biometrika. 3, 631-641.

Kovoor, G.M., and Nandagiri, L. 2007. Developing regression models for predicting pan evaporation from climate data. Journal of Irrigation and Drainage Engineering. 133, 444- 454.

Ladlani, I., Hauichi, L., Dhemili, L., Heddem, S., and Blouze, K.h. 2013. Estimation of Daily Reference Evapotranspiration in the North of Algeria using Adaptive Neuro-Fuzzy Inference System (ANFIS) and Multiple Linear Regression (MLR) Models: A Comparative Study. Arabian Journal for Science and Engineering. 39, 5959-5969.

Liu, C.W., Lin, K.H., and Kuo, Y.M. 2003. Application of factor analysis in the assessment of groundwater quality in a blackfoot disease area in Taiwan. Journal of Science of the Total Environment. 313, 77-89.

Maidement, R. 1993. Handbook of Hydrology: New York: Mc GRAW-Hill, INC.

Malik, A., and Kumar, A. 2015. Pan evaporation simulation based on daily meteorological data using soft computing techniques and multiple minear megression. Water Resources Management. 29, 1859-1872.

Malik, A.P., Jyothy, S.A., and Sekhar Reddy, K.C. 2013. Daily Reference Evapotranspiration Estimation using Linear Regression and ANN Models. The Institution of Engineers (India). 93(4), 215–221.

Nazemosadat, S.M.J., and Shirvani, A. 2005. Prediction of Persian Gulf Sea Surface Temperature Using Multiple Regressions and Principal Components Analysis. Journal of Water and Soil Science. 9 (3), 1- 11. (In Persian)

Neal, R.A., and Phillips, I.D. 2009. Summer daily precipitation variability over the East Anglian

region of Great Britain. Journal of Climatology. 29, 1661-1679.

Noori Gheidari, M.H. 2010. Identify outliers in regional flood analysis using principal component analysis. Proceeding of 5th National Congress on Civil Engineering in Mashhad University, Mashhad, Iran. May 4- 6. (In Persian)

Raziei, T., and Azizi, G. 2009. Delineation of homogeneous precipitation regions in Western Iran. Journal of Geography and Environmental Planning. 20 (2), 65-86. (In Persian)

Salas, J.D. 1993. Analysis and Modeling of Hydrological Time Series: Handbook of Hydrology (edited by David R. Maidment): New York: McGraw-Hill.

Seifi, A., Mirlatifi, S.M., and Riahi, H. 2011. Developing a Combined Model of Multiple Linear Regression-Principal Component and Factor Analysis (MLR-PCA) for Estimation of Reference Evapotranspiration (Case Study: Kerman Station). Journal of Water and Soil. 24 (6), 1186- 1196. (In Persian)

Sheikholeslami, N., Ghahraman, B., Mosaedi, A., Davary. K., and Mohejerpour, M. 2014. Estimating Reference Evapotranspiration by Using Principal Component Analysis (PCA) and The Development of a Regression Model (MLR-PCA) (Case Study: Mashhad Station). Journal of Water and Soil. 28(2), 420- 429. (In Persian)

Shirgure, P.S. 2011. Evaporation modeling with neural networks-A research review. Int. J. Res. Rev Soft Intell. Comput. 1(2), 37–47.

Shirgure, P.S., and Rajput, G.S. 2012. Prediction of daily pan evaporation using neural networks models. Science Journal Agriculture. 1(5),126–137.

Shirsath, P.B., and Kumar, A.S. 2009. A comparative study of daily pan evaporation estimation using ANN, regression and climate based models. Water Resources Management. 24, 1571-1581.

Shirsath, P.B., and Singh, A.K. 2010. A comparative study of daily pan evaporation estimation using ANN, regression and climate based models. Water Resources Management. 24,1571–1581.

Shirvani, A., Nazemosadat, S.M.J. 2012. Regionalization of precipitation in Iran using principal components and cluster analysis. Journal of Iran-Water Resources Research. 8(1), 81- 85. (In Persian)

Shrestha, S., and Kazama, F. 2007. Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan. Journal of Environment Modelling Software. 22, 464- 475.

Tianxiao, L., Qiang, F., SHugin, X., and Fanxiang, M. 2009. Application of principal component analysis in evaluating influence factors of evaporation in northern cold area. Proceeding of Fifth International Conference on Natural Computation. Tianjin. Augest 14- 16.

Yu, P.S.H., and Yung, T.C.H. 1996. Synthetic regional flow duration curve for Southern Taiwan. Journal of Hydrological Processes. 10, 373-391.