Applied Water Science (2020) 10:130
https://doi.org/10.1007/s13201-020-01212-4

ORIGINAL ARTICLE

Fuzzy nearest neighbor approach for drought monitoring and assessment

E. Fadaei-Kermani[1] · M. Ghaeini-Hessaroeyeh[1]

Received: 10 March 2020 / Accepted: 28 April 2020 / Published online: 12 May 2020

Abstract

In this paper, a new approach is proposed based on the *Fuzzy*-nearest neighbor model to deal with drought monitoring. According to the Standardized Precipitation Index and via *Fuzzy*-kNN approach, a method has been presented to predict the most likely drought conditions. In order to appraise the precision of results, the model was applied to monitor the drought status in city of Kerman, located in south east of Iran. The results showed that the area has faced drought and also rainfall shortages in recent years. The calculated values of correlation coefficient, RMSE, CRM and MAE coefficients showed the accuracy and efficiency of the proposed approach.

Keywords *Fuzzy*-kNN model · SPI · Drought monitoring · Drought Index · Model evaluation

Introduction

Drought is a natural and repetitive phenomenon which occurs due to reduction in rainfall over a certain period. That whether or not this phenomenon can be short and less severe depends on its extremity, continuance, and the distance of impacted area. It may start slowly but emerge in a relatively long interval in different sections including agriculture, water resources, economy, environment, etc. (Mishra and Singh 2010). Drought can occur in any climatic conditions throughout the world. The designing and management of water resources and different agricultural sections are highly related to how manage drought and adopt proper guidelines to face such phenomena (Fadaei-Kermani et al. 2017).

Drought prediction and monitoring can play a very important role in the system management of water resources and remarkably decrease the damage. In general, the intensity of drought is predicted and monitored via drought indices. The drought indices aim to expound the phenomenon quantitatively and also include the combination of different effective features on drought in quantitative and simple relations. There are usually various indices to monitor this phenomenon including: Palmer-Drought Severity Index (PDSI) (Palmer 1968), Deciles Index (DI) (Gibbs and Maher 1976), Standardized Precipitation Index (SPI) (McKee et al. 1993), Reclamation-Drought Index (RDI) (Weghorst 1996), US Drought Monitor (USDM) (Svoboda et al. 2002) and, etc.

In recent years, drought and its dependent crises and threats have become one of the most important global challenges. A large number of research has been conducted regarding the drought monitoring and control techniques (e.g., Luo and Wood 2007; Paulo and Pereira 2008; Rhee et al. 2010; Pan et al. 2013; Fadaei-Kermani et al. 2014; Hao and AghaKouchak 2014; Wood et al. 2015; Hao et al. 2016; Park et al. 2017; Yu et al. 2018; Abbasi et al. 2019). These studies used drought indices, machine learning, and data mining algorithms to monitor and predict the severity of the effects caused by drought.

By average annual rainfall of 240–250 mm as one third of average global figure, Iran is considered among the regions in which these are insufficient proper precipitation. Since most parts of this country are covered with dried areas, water has played a vital role in its economic development. In the present study, via the fuzzy *k*-nearest neighbor model, a method has been proposed to predict the most likely drought status of Kerman, south eastern of Iran. The nonparametric techniques (e.g., fuzzy-*k*-nearest neighbor algorithm) can be applied as convenient approaches for estimating drought conditions. These algorithms can be useful in problems that

E. Fadaei-Kermani
ehsanhard@gmail.com

[1] Department of Civil Engineering, Faculty of Engineering, Shahid Bahonar University of Kerman, PO Box 76169-133, Kerman, Iran

the relationships between instances are not already obvious and fully determined.

Standard Precipitation Index (SPI)

McKee et al. (1993) proposed the SPI (Standardized Precipitation Index) for drought monitoring respect to multiple time scales. SPI is widely used for characterizing and detecting meteorological drought which can be compared across regions with significantly different climate. It is determined using long-term precipitation records, and then a Z-standard normal distribution is fitted according to following equations:

$$\mathrm{SPI} = -\left[a - \frac{E_0 + E_1 a + E_2 a^2}{1 + d_1 a + d_2 a^2 + d_3 a^3}\right] \qquad a = \sqrt{\ln\left[\frac{1}{P(x)^2}\right]} \quad 0 < P(x) \le 0.5 \tag{1}$$

$$\mathrm{SPI} = +\left[a - \frac{E_0 + E_1 a + E_2 a^2}{1 + d_1 a + d_2 a^2 + d a^3}\right] \qquad a = \sqrt{\ln\left[\frac{1}{1 - P(x)^2}\right]} \quad 0.5 < P(x) \le 1 \tag{2}$$

$$\begin{aligned} d_1 &= 1.432788 \quad & E_0 &= 2.51557 \\ d_2 &= 0.189269 \quad & E_1 &= 0.802853 \\ d_3 &= 0.001308 \quad & E_2 &= 0.010328 \end{aligned}$$

where $P(x)$ refers to the cumulative probability function. According to precipitation data, time series can be obtained. After the data were sorted in increasing order, the empirical probability distribution is calculated as follows (Fadaei-Kermani et al. 2017):

$$\mathrm{ECP} = \frac{a}{b+1} \tag{3}$$

where a equals the row number of sorted precipitation data, and b represents the precipitation data total number. The standard normal cumulative distribution curves can be used to calculate the Standard Precipitation Index (SPI) for each corresponding time scales.

Table 1 represents the drought intensities classification according to the range of SPI values. Anytime the SPI value is continuously negative, drought is likely to occur. On the other hand the event ends when the SPI value becomes positive (Moreira et al. 2006).

Table 1 Classification of drought intensity due to SPI values

Class	Range of SPI values	Status of drought
I	+2 and more	Extremely wet
II	1.5 to 1.99	Very wet
III	1 to 1.49	Moderately wet
IV	0.99 to −0.99	Near normal
V	−1.49 to −1	Moderately dry
VI	−1.99 to −1.5	Very dry
VII	−2 and less	Extremely dry

Fuzzy-nearest neighbor algorithm

As one of the most popular nonparametric and lazy instance based machine learning algorithms, the k-nearest neighbor algorithm is extensively applied in data mining and pattern recognition (Fadaei-Kermani et al. 2015). In recent years, several approaches to nearest neighbor modeling have been suggested based on fuzzy mathematics to improve the quality of the classification. Keller et al. (1985) proposed a fuzzy version of the basic k-NN algorithm by incorporating the theory fuzzy sets into the standard k-NN. It was named *Fuzzy*-nearest neighbor algorithm (*Fuzzy*-kNN). The both k-nearest neighbor and *Fuzzy*-nearest neighbor algorithms involve measuring the similarity of a new instance (unknown instance) to the instances with a specific label in the training set. Then, by determining a set of k nearest neighbors, and casting a vote on the class of query instances, the most likely class can be dedicated to the unknown instance by incorporating all the votes (Derrac et al. 2016 and Ezghari et al. 2017).

Owing to the *Fuzzy*-nearest neighbor algorithm, rather than individual classes as in the k- nearest neighbor modeling, a Fuzzy Membership Function (FMF) of samples can be specified to all various categories (Kermani et al. 2018). Let $X = \left(x_1, x_2, \ldots, x_n\right)$ be a training set consists of n labeled samples which are introduced by C classes. In case of a new unknown instance Y, the class confidence values can be determined as the aggregation of k nearest neighbors' class attributes according to Eq. 4 (Keller et al. 1985).

$$u_i(y) = \frac{\sum_{j=1}^{k} \mu_{ij}\left(1/\|y - x_j\|^{2/(m-1)}\right)}{\sum_{j=1}^{k}\left(1/\|y - x_j\|^{2/(m-1)}\right)} \tag{4}$$

where $i = 1, 2, \ldots, C$, and $j = 1, 2, \ldots, k$. The fuzzy strength parameter m is utilized to intensify the distances between the unknown instances and the related elements of training data set. The value of m can be chosen as $m \in (1, +\infty)$ that is often $m = 2$. $y - x_j$ expresses the distance between y and its jth nearest neighbor from the training set data x_j. μ_{ij} refers to the membership rating of the instance x_j among the training set to the class i, among the k nearest neighbors of x that is satisfied the following relations (Derrac et al. 2016):

$$\mu_{ij} \in [0,1] \tag{5a}$$

$$0 < \sum_{j=1}^{k} \mu_{ij} < k \tag{5b}$$

$$\sum_{i=1}^{c} \mu_{ij} = 1 \tag{5c}$$

where $1 \le i \le C$ and $1 \le j \le k$.

In the general *Fuzzy*-kNN model, various techniques can be applied to define μ_{ij}. In the case of crisp labeling, every instance has membership of one in its known class and zero-membership in other classes. In case of a constrained fuzzy membership, the k nearest neighbors of every training set data (x_k) is determined, and then the membership of x_k in every class can be calculated using the following membership function (Keller et al. 1985):

$$\mu_{ij}\left(x_k\right) = \begin{cases} 0.51 + \left(\frac{n_j}{K}\right) * 0.49, & \text{if } j = i \\ \left(\frac{n_j}{K}\right) * 0.49, & \text{otherwise} \end{cases} \tag{6}$$

where n_j represents the neighbors number found which fit in the jth class. The fuzzy procedure causes no arbitrary assignments can be made by the algorithm. Moreover, a level of assurance should be provided by the membership values of the vector to attend the outcome classification.

Model processing and application

In the present study, the hydrological and precipitation data of Kerman city during 1980–2018 has been investigated. The area is located in southeast of Iran between 53° and 26 min to 59° and 29 min of eastern length and 25° and 55 min to 32° northern latitudes (Fig. 1). Drought has been always a prevalent phenomenon in Kerman province. The area has never been detached from the destructive consequences of this phenomenon.

According to the precipitation data of Kerman city, the moving time series and, respectively, the standard normal distribution functions were determined based on different time scales. By calculating the standard normal cumulative distribution, the SPI value can be obtained for every corresponding time scales. Figure 2 presents the precipitation cumulative and standard normal probability distribution functions of the Kerman precipitation data for 3-, 6-, 12-, 24- and 48-month time scales. These graphs can be used to determine the SPI value and corresponding drought status according to the precipitation data for every time scales.

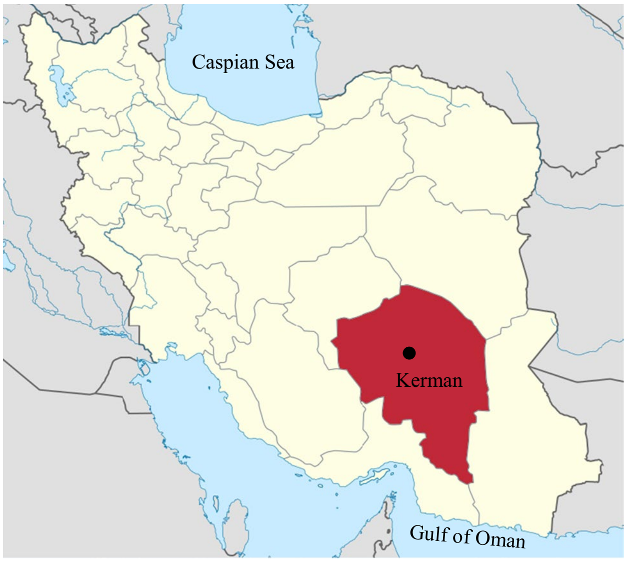

Fig. 1 The location of Kerman province in Iran map

Then, according to the standard normal probability and the precipitation cumulative probability distribution functions, the SPI values for various time scales have been calculated. For example, Fig. 3 shows the calculated values of 3-, 6-, 12- and 24-month SPI for the study area during different years.

Then the calculated values of SPI during the desired period can be applied in the *Fuzzy*-nearest neighbor model. Before working with the model, the data should be normalized using the relation 7.

$$Y' = \frac{y - \overline{y}}{\sigma(y)} \tag{7}$$

where the normalized variable value (Y') can be obtained according to standard deviation ($\sigma(y)$) and mean ($\overline{y}$) of the observed variable values in the reference dataset.

Finally, the accuracy and efficiency of the model can be evaluated via root-mean-square error (RMSE), mean absolute error (MAE), coefficient of correlation (r) and coefficient of residual mass (CMR). These coefficients can be obtained using following equations (Dashtaki et al. 2009):

$$\text{MAE} = \frac{\sum_{i=1}^{n} \left|x_i - y_i\right|}{n} \tag{8}$$

$$r = \frac{n\left[\sum_{i=1}^{n} y_i x_i\right] - \left[\sum_{i=1}^{n} y_i\right]\left[\sum_{i=1}^{n} x_i\right]}{\sqrt{\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]}} \tag{9}$$

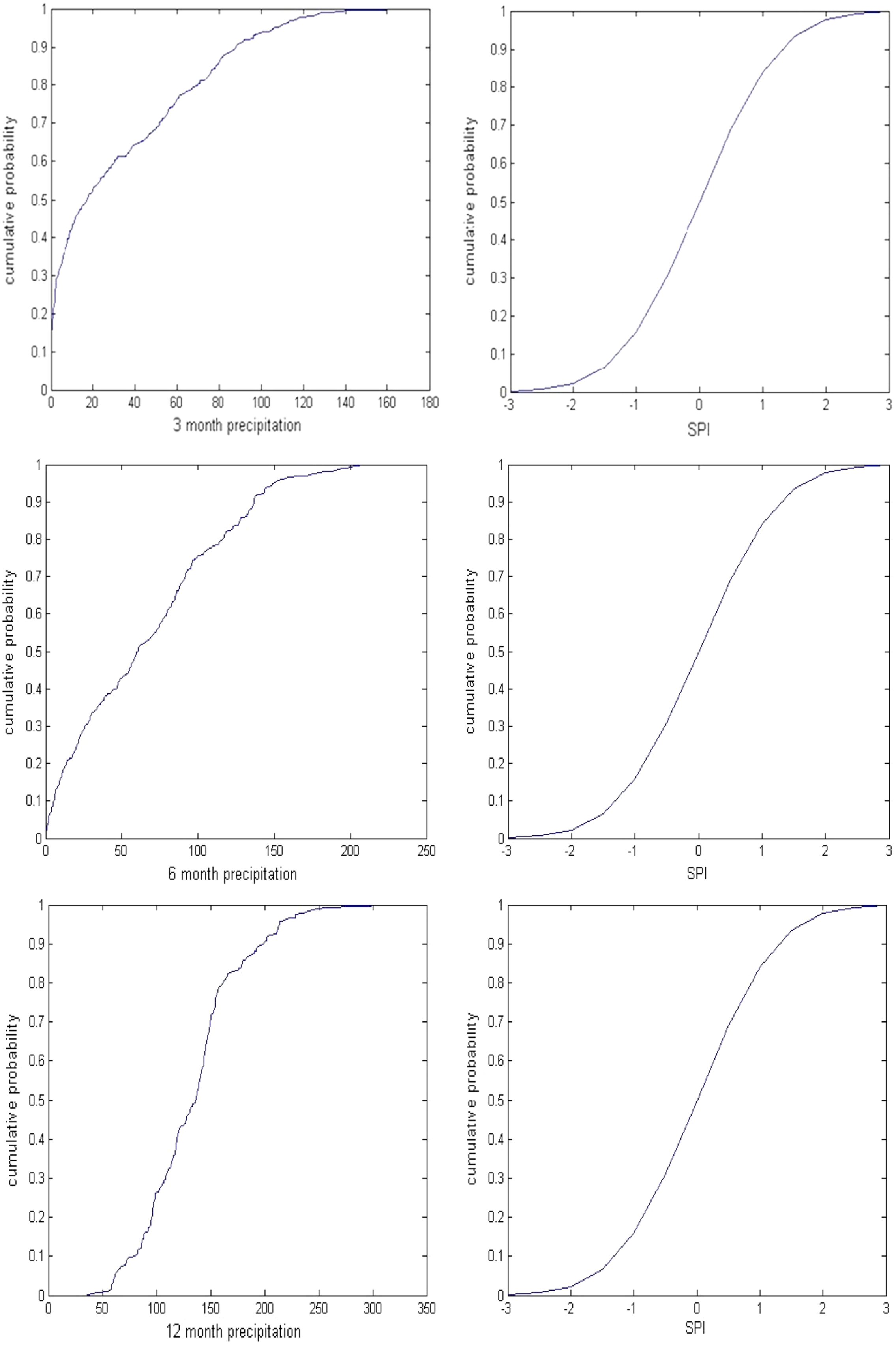

Fig. 2 The precipitation cumulative and standard normal probability distribution functions

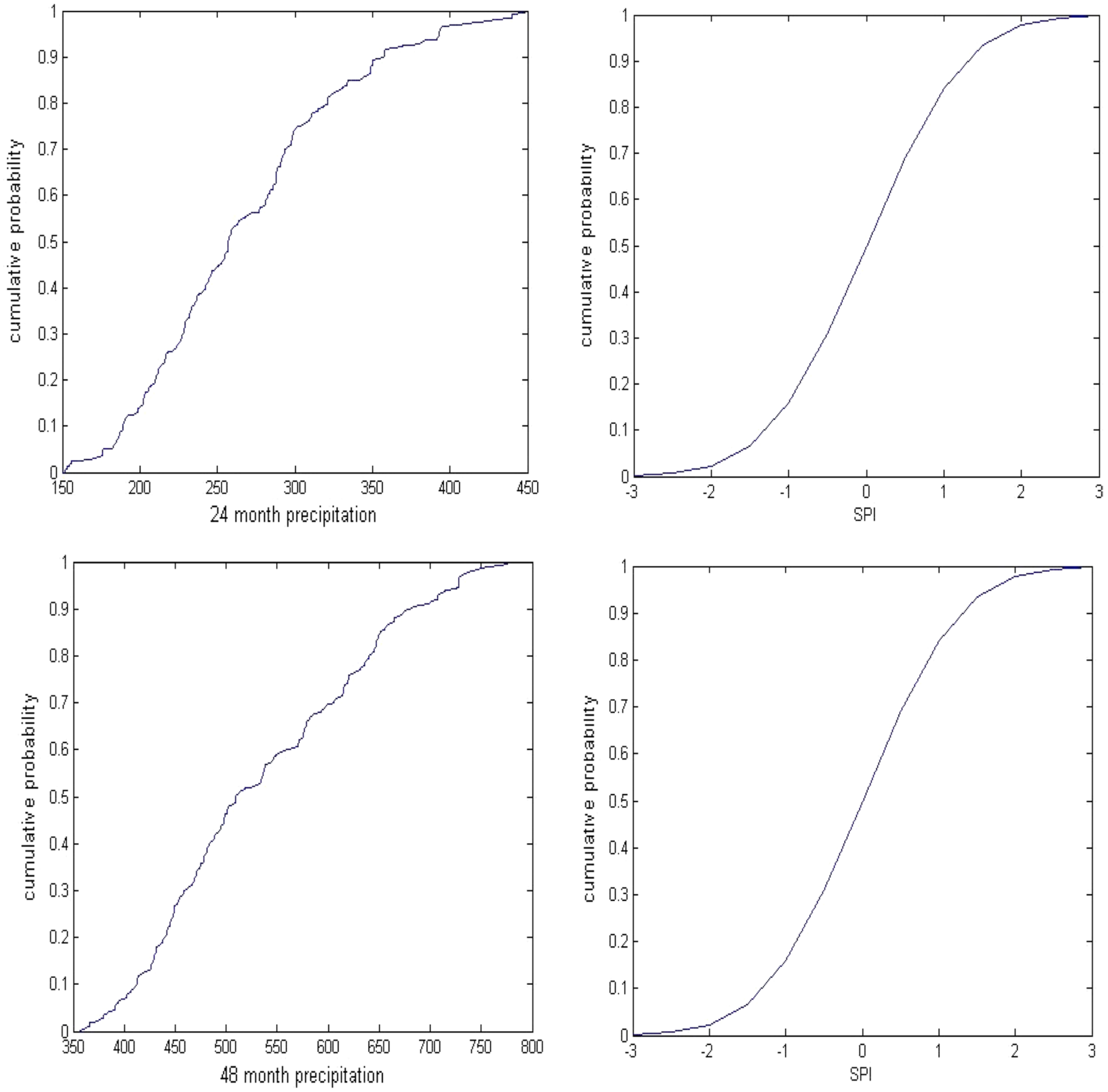

Fig. 2 (continued)

$$\mathrm{RMSE} = \left[\frac{\sum_{i=1}^{n} (x_i - y_i)^2}{n} \right]^{0.5} \tag{10}$$

$$\mathrm{CRM} = \frac{\left(\sum_{i=1}^{n} x_i\right) - \left(\sum_{i=1}^{n} y_i\right)}{\sum_{i=1}^{n} x_i} \tag{11}$$

where y_i and x_i express the values of predicted and measured attributes, and n refers to the number of attributes.

Results and discussion

At the beginning of the calculations, the number of nearest neighbors of the attributes for the *Fuzzy*-kNN model should be determined. The best value of K (number of nearest neighbor) can be determined by n-fold cross-validation method. First, the data set is divided into n equal-sized parts (Fig. 4). For each part, the model is trained to the other data set parts, and the prediction error of the fitted model is calculated when the desired part of the data is predicted. The procedure is done for every value of k ($k = 1, 2, \ldots, K$) to obtain the best value of k with minimum prediction error rate (Huang et al. 2017). In Fig. 5 the precision of the fourfold cross-validation method according to the sum of squares error (SSE) coefficient has been shown. Due to

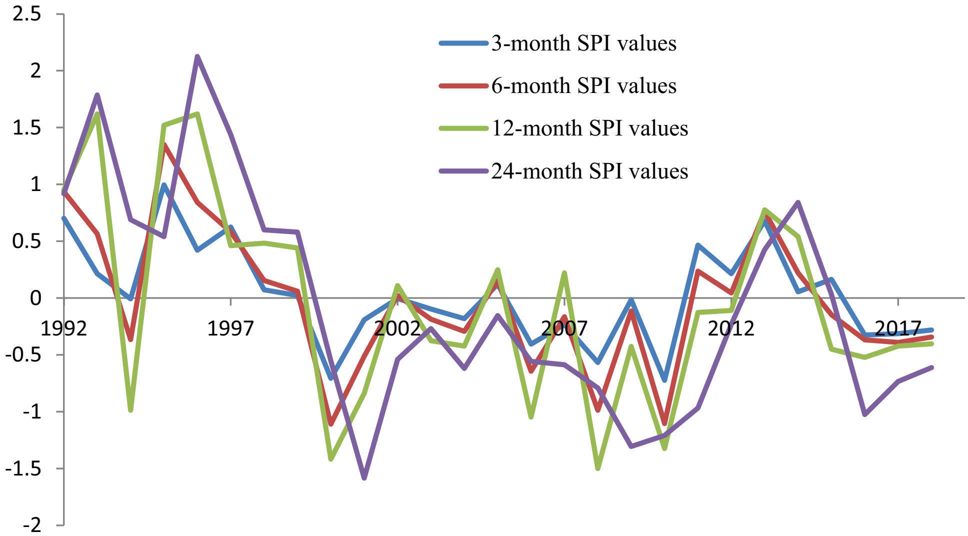

Fig. 3 The 3-, 6-, 12- and 24-month SPI values for different years

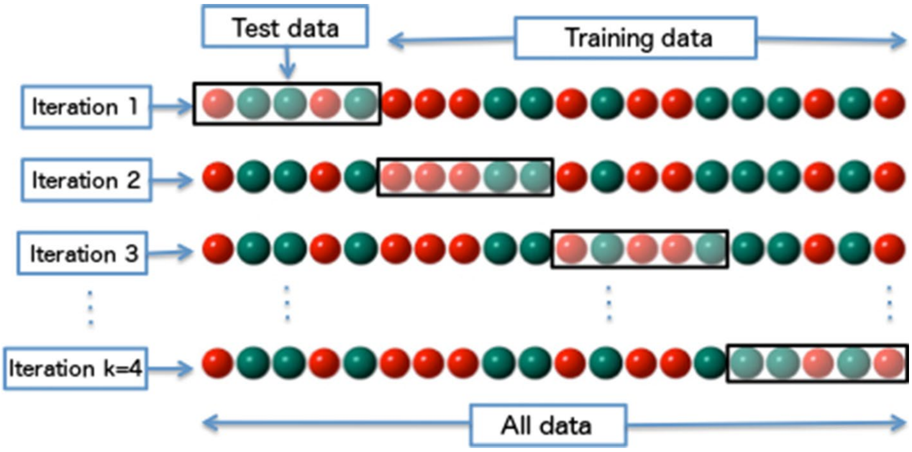

Fig. 4 The *n*-fold cross-validation method scheme

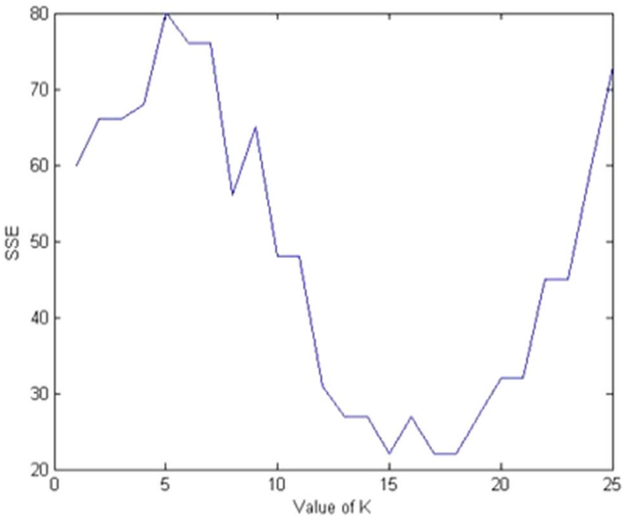

Fig. 5 The error rate for *Fuzzy*-kNN model according to fourfold-cross-validation

Fig. 5, the values of $k = 15$, 17 and 18 have the same lowest error rate. The K value of 18 has been selected for the *Fuzzy*-kNN model since according to Travis and Mays (2010) the larger k values can often minimize risk of overfitting.

After determining the best value of K, the value can be introduced to the *Fuzzy*-kNN model for further computations. Then according to calculated SPI values, the most likely drought situation for the city of Kerman was determined during different years. Table 2 and Fig. 6 present the region drought classification determined and predicted by the *Fuzzy*-kNN model.

According to the results, Kerman has recently been exposed to drought and also rainfall shortages in normal and even much lower than normal levels. This is clearly evident in the drought classes which are assigned to the region (classes 4 and above). Since the average annual precipitation in Kerman is about 122 mm compared to the average annual rainfall of Iran (about 250 mm), which is very low on the global scale, it

Table 2 Drought classification assigned by the *Fuzzy*-kNN model for the city of Kerman

Assigned membership to drought classes	Year						
	Class-I	Class-II	Class-III	Class-IV	Class-V	Class-VI	Class-VII
2010	0.0000	0.0000	0.0000	0.0000	0.0000	0.1783	0.8217
2011	0.0000	0.0000	0.0986	0.9014	0.0000	0.0000	0.0000
2012	0.0000	0.0000	0.2870	0.7130	0.0000	0.0000	0.0000
2013	0.0000	0.0000	0.0534	0.8534	0.0932	0.0000	0.0000
2014	0.0000	0.0000	0.0000	0.9812	0.0188	0.0000	0.0000
2015	0.0000	0.0000	0.0725	0.9275	0.0000	0.0000	0.0000
2016	0.0000	0.0000	0.0000	0.0000	0.1275	0.7860	0.0865
2017	0.0000	0.0000	0.0000	0.3927	0.3955	0.2118	0.0000
2018	0.0000	0.0000	0.0000	0.5307	0.3554	0.1139	0.0000

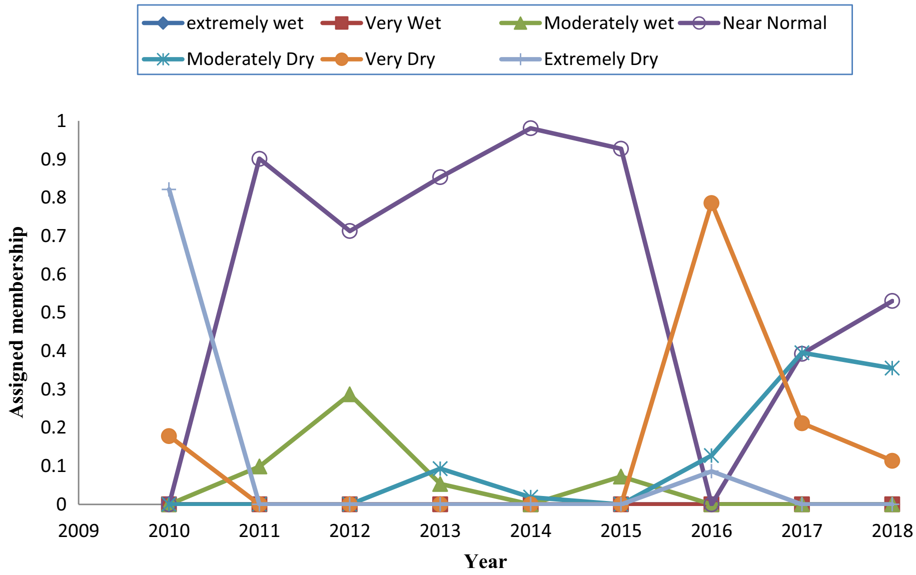

Fig. 6 Assigned membership to each drought class according to the *Fuzzy*-kNN model

Table 3 The *Fuzzy*-kNN model evaluation

r	RMSE	CRM	MAE
0.924	0.108	0.0012	0.101

indicates the damage and the intensity of the phenomenon in this region. The accuracy and precision of the present model results have been evaluated by coefficient of correlation (r), root-mean-square error (RMSE), coefficient of residual mass (CRM) and mean absolute error (MAE). The calculated values are presented in Table 3. Appropriate values of the coefficients indicate the acceptable precision and low error rate of the *Fuzzy*-kNN model for drought monitoring and prediction.

Conclusion

Drought is a climatic phenomenon, occurring in any climatic conditions which can affect different aspects of water resources management and planning. The present study deals with the investigation of drought intensity and status in city of Kerman located in Iran. In this paper, a new approach was proposed to monitor and predict the most likely drought status for the study area via the *Fuzzy*-kNN model. At first according to the precipitation data, the values of SPI for different time scales were determined. Then the *Fuzzy*-kNN modeling was employed to predict the most likely drought status. The results showed that this city has faced drought and also rainfall shortages in the recent years which are consistent with real observations. Finally, the values of coefficient of correlation ($r = 0.924$), root-mean-square error (RMSE = 0.108), coefficient of residual mass (CRM = 0.0012) and mean absolute error (MAE = 0.101) were calculated according to the results of *Fuzzy*-kNN modeling. The results indicated that the present model is efficient and accurate.

Compliance with ethical standards

Conflict of interest The authors declare that there is no conflict of interests regarding the publication of this paper.

References

Abbasi A, Khalili K, Behmanesh J, Shirzad A (2019) Drought monitoring and prediction using SPEI index and gene expression programming model in the west of Urmia Lake. Theor Appl Climatol 138:553–567

Dashtaki SG, Homaee M, Mahdian MH (2009) Site-dependence performance of infiltration models. Water Resour Manag 23:2777–2790

Derrac J, Chiclana F, García S, Herrera F (2016) Evolutionary fuzzy k-nearest neighbor algorithm using interval-valued fuzzy sets. Inf Sci 329:144–163

Ezghari S, Zahi A, Zenkouar K (2017) A new nearest neighbor classification method based on fuzzy set theory and aggregation operators. Expert Syst Appl 80:58–74

Fadaei-Kermani E, Barani GA, Ghaeini-Hessaroeyeh M (2015) Prediction of cavitation damage on spillway using K-nearest neighbor modeling. Water Sci Technol 71(3):347–352

Fadaei-Kermani E, Barani GA, Ghaeini-Hessaroeyeh M (2017) Drought monitoring and prediction using K-nearest neighbor algorithm. J AI Data Min 5(2):319–325

Gibbs WJ, Maher JV (1976) Rainfall deciles as drought indicates. Australian Bureau of Meteorology, Bull, pp 37–48

Hao Z, AghaKouchak A (2014) A nonparametric multivariate multi-index drought monitoring framework. J Hydrometeorol 15(1):89–101

Hao Z, Hao F, Singh VP (2016) A general framework for multivariate multi-index drought prediction based on Multivariate Ensemble Streamflow Prediction (MESP). J Hydrol 539:1–10

Huang J, Keung JW, Sarro F, Li YF, Yu YT, Chan WK, Sun H (2017) Cross-validation based K nearest neighbor imputation for software quality datasets: an empirical study. J Syst Softw 132:226–252

Keller M, Gray MR, Givens JA (1985) A fuzzy k-nearest neighbor algorithm. IEEE Trans Syst Man Cybern 15:580–585

Kermani EF, Abbas Barani G, Javad Khanjani M (2014) Developing a framework for compatibility analysis of predictive climatic variables distribution with reference evapotranspiration in probabilistic analysis of water requirement. J Appl Res Water Wastewater 1(2):66–73

Kermani EF, Barani GA, Hessaroeyeh MG (2018) Cavitation damage prediction on dam spillways using Fuzzy-KNN modeling. J Appl Fluid Mech 11(2):323–329

Luo L, Wood EF (2007) Monitoring and predicting the 2007 US drought. Geophys Res Lett 34(22):L22702

McKee TB, Doesken NJ, Kleis J (1993) The relationship of drought frequency and duration to time scales. In: Eighth conference on applied climatology, 17–22 January, Anaheim, California

Mishra AK, Singh VP (2010) A review of drought concepts. J Hydrol 391:202–216

Moreira EE, Paulo AA, Pereira LS, Mexia JT (2006) Analysis of SPI drought class transitions using loglinear models. J Hydrol 331(1–2):349–359

Palmer WC (1968) Keeping track of crop moisture conditions, nationwide: the new crop moisture index. Weather-wise 21:156–161

Pan M, Yuan X, Wood EF (2013) A probabilistic framework for assessing drought recovery. Geophys Res Lett 40(14):3637–3642

Park S, Im J, Park S, Rhee J (2017) Drought monitoring using high resolution soil moisture through multi-sensor satellite data fusion over the Korean peninsula. Agric For Meteorol 237:257–269

Paulo AA, Pereira LS (2008) Stochastic prediction of drought class transitions. Water Resour Manag 22(9):1277–1296

Rhee J, Im J, Carbone GJ (2010) Monitoring agricultural drought for arid and humid regions using multi-sensor remote sensing data. Remote Sens Environ 114(12):2875–2887

Svoboda MD, Le Comte D, Hayes MJ (2002) The drought monitor. Bull Am Meteorol Soc 93(8):1181–1190

Travis QB, Mays LW (2010) Prediction of intake vortex risk by nearest neighbors modeling. J Hydraul Eng 137(6):701–705

Weghorst KM (1996) The Reclamation Drought Index: guidelines and practical applications. Bureau of Reclamation, Denver, CO, p 6 (Available from Bureau of Reclamation, D-8530, Box 25007, Lakewood, CO 80226)

Wood EF, Schubert SD, Wood AW, Peters-Lidard CD, Mo KC, Mariotti A, Pulwarty RS (2015) Prospects for advancing drought understanding, monitoring, and prediction. J Hydrometeorol 16(4):1636–1657

Yu J, Lim J, Lee KS (2018) Investigation of drought-vulnerable regions in North Korea using remote sensing and cloud computing climate data. Environ Monit Assess 190(3):126

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.