PREDICTION OF ANNUAL RUNOFF AT THE DANJIANGKOU RESERVOIR, CHINA BASED ON FORECAST DOMAIN

YANG, M. X.[1*] – ZHANG, Y.[2*] – WANG, H.[1] – JIANG, Y. Z.[1] – XU, Z.[3] – LEI, X. H.[1]

[1]*State Key Laboratory of Simulation and Regulation of Water Cycle in River Basin, China Institute of Water Resources and Hydropower Research, Beijing 100038, China*

[2]*School of Software, Nanchang Hangkong University, Nanchang 330063, China*

[3]*China Water Rights Exchange Company Limited, Beijing 100053, China*

**Corresponding authors*
e-mail: yangmx@iwhr.com; phone:+86-18-046-555-306
e-mail: zyan_iwhr@163.com; phone: +86-13-651-210-838

(Received 27th Nov 2018; accepted 1st May 2019)

Abstract. As the water resource management progresses rapidly in recent years, middle and long-term runoff forecast has become increasingly important. Conventional multi-category runoff prediction usually utilizes manually specified threshold values to categorize runoff categories. However, this approach is arguably subjective, and it neglects fuzziness and peculiarity of hydrometeorological time series. To address this issue, a new concept, forecast domain, is proposed in this study. Cluster analysis of runoff time series was carried out with the Gaussian Mixture Model, and Support Vector Classification was then used to establish the nonlinear relationships between forecast domain and various potential predictors. The current study focuses on the Danjiangkou Reservoir, the source of the Central Route of the South-North Water Transfer Project in China. We use the 25-year data (1981-2005) for model training, and the Danjiangkou runoff data during last 11 years (2006-2016) are used for model validation. It is shown that the runoff forecast domain obtained from the unsupervised clustering is reasonable and appropriate for categorizing runoff categories. Further forecast experiments reveal that this model may shed some light on the prediction of annual mean runoff at the Danjiangkou Reservoir.
Keywords: *middle and long-term runoff forecast, category runoff prediction, Gaussian Mixture Model, cluster analysis, Support Vector Classification*

Introduction

Middle and long-term runoff forecast refers to the quantitative or qualitative prediction of runoff of various water bodies, e.g., rivers, reservoirs, and lakes, over the time horizon from more than three days up to one year. Runoff forecast is usually based on the past and present hydrological and meteorological information, as well as fundamental principles and methodologies based on several related disciplines including hydrology, meteorology, hydrodynamics, and statistics (Fan, 1999). As the water resource management progresses rapidly in recent years, middle and long-term runoff forecast has become increasingly more important and pressing.

Recent studies have made effort to address this issue. Yang et al. (2005) constructed a long-term runoff forecast system during the dry season by combining the continuous rainfall-runoff model and the long-term weather outlook. Their model was found to perform reasonably well. Hong et al. (2016) applied the genetic algorithm to improve the phase-space reconstruction method, and developed a new nonlinear model for monthly mean runoff. Their model was tested in four types of experiments using data from six hydrological stations along the Yellow River and the Yangtze River. Forecast experiments

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1704_95619575

show that the medium- and long-term runoff forecast is satisfactory at these stations. A number of other recent studies further applied recently developed machine learning and artificial intelligence techniques to extended range runoff forecasts. Coulibaly et al. (2015) investigated the impact of climate trends on the forecast accuracy using a recurrent neural network (RNN), which was trained using time series of runoff to eight large hydropower systems in Quebec and Labrador and several selected climate indices. Results from the forecast experiments indicate that the use of BWA, PNA and ENSO indices leads to better forecast skill than the SLP or NAO indices alone. Maslova et al. (2016) constructed a model by combing wavelet decomposition and Bayesian machine learning regression techniques. The authors compared their model with that of the wavelet and artificial neural networks-based model and evaluated the effects of different wavelet boundary rules with synthetic and real runoff data collected from the Yellowstone River in the Uinta Basin in Utah. It is shown that their model accuracy can be improved by using a new wavelet boundary rule introduced in that study. Yang et al. (2017) applied three machine learning techniques to runoff forecast: Random Forest (RF), Artificial Neural Network (ANN) and Support Vector Regression (SVR). They compared the performance for forecasting one-month-lead reservoir runoff for two headwater reservoirs in USA and China, respectively. It is shown that RF yields the best statistical performances among the three. Tan et al. (2018) made an attempt to improve the decomposition-ensemble framework and proposed an adaptive model for medium and long-term runoff forecast in both the dry and flood seasons. The authors recommended to use SAR (1) model in the dry season and AEEMD-ANN model in the flood season to forecast the monthly runoff in Yangtze River Basin.

Formation of runoff is a result of complex interaction among a range of physical processes, including precipitation, evaporation, and confluence, and human activities. The category of complexity dictates that runoff forecast is inherently stochastic and highly uncertain. Therefore, it is challenging to predict the accurate value of future runoff based on qualitative analysis of physical processes. Besides, one may predict runoff categories instead of single values. The extension of prediction from single values to the runoff categories may help improve forecast reliability and enhance practical values of runoff forecast for the development and utilization of water resources. Indeed, past experience suggests that predicting runoff categories is more reliable and informative compared to single value prediction, as it improves the precision of the forecast e.g. Kasiviswanathan et al. (2013) and Ye et al. (2014). Quan et al. (2014) developed a method to construct prediction rainfall runoff categories with an artificial neural network (ANN) model. The model was calibrated by generating ensemble predictions, and tested in a real-world case study of rainfall-runoff data. The authors showed that the peak flows are predicted with improved accuracy with this method compared to traditional single point forecasts by ANNs. Li et al. (2017) calibrated and validated the different distribution types of Bayesian forecasting system for the observed 52 floods during 2004-2014 at the ZheXi basin. They showed that the Log Weibull and empirical Bayesian probabilistic model perform the best on average compared with the other distribution models. However, one limitation in these studies is that these methods require manually specifying the prediction categories. Because of the fuzziness of runoff, it may be argued that these conventional methods neglect peculiarity of hydrometeorological time series.

One method to objectively classify runoff categories is to apply sequence clustering analysis, which takes full consideration of fuzziness in runoff time series and provides faithful representation of the physical laws governing runoff. For example, Hou et al. (2016) combined three methods: sequenced sample clustering, set pair analysis (SPA), and Markov

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1704_95619575

chains; this approach leads to multiple improvements compared to the conventional weighted Markov chains. Based on these methods, the authors constructed a prediction model for annual mean precipitation. The results show that the improvements lead to better classification of precipitation, sharper forecast probability distribution, and improved forecast precision. Zhao et al. (2017) proposed to use cluster analysis to examine anomaly correlation, a performance measure of raw general circulation model forecasts in the three-dimensional space of latitude, longitude, and initialization forecast time. Totz et al. (2017) developed a new cluster-based empirical model to forecast winter precipitation anomalies. They compared this model with dynamic forecast models and a canonical correlation analysis-based prediction model. The results indicated that this new prediction method performs better regarding timing and pattern correlation in the Mediterranean and European regions. Another widely used clustering method is the Gaussian Mixture Model (GMM), which is a parametric model based on Gaussian distribution and trained with the Expectation Maximization algorithm. The GMM can objectively classify runoff without human intervention, hence better suitable for classifying runoff categories for prediction.

Motivated by these efforts and considerations, we proposed the concept of forecast domain. Cluster analysis was conducted on the multiple year runoff data using the GMM, and resultant clusters are then used to construct forecast domain. Finally, Support Vector Classification (SVC) is adopted to predict the forecast domain. We apply this model to the forecast runoff at the Danjiangkou Reservoir, which is the primary water source of the Central Route of the South-North Water Transfer Project.

This paper make two main contributions. On the one hand it applied sequence clustering analysis which takes full consideration of fuzziness in runoff time series and provides faithful representation of the physical laws governing runoff. On the other hand it proposed a concept of forecast domain to expand the prediction results from specific values to the range, and the characteristics of the runoff change interval state were more evident.

Materials and Methods

Overview of the geography and climate of the Danjiangkou Reservoir area

The Danjiangkou (DJK) Reservoir (110 °E - 112°E，32°N-33°N, abbreviated as DJK hereafter) is the largest artificial fresh water lake in Asia (Li and Zhang, 2014). DJK is located at the boundary between the Danjiangkou city, Hubei Province and Zhechuan County, Hunan Province. DJK is the confluence of Han River and Danjiang River, with a drainage area of 17,916 km^2. The terrain of DJK is characterized by great elevation differences, steep slopes, and deeply dissected topography. The highest altitude reaches 1,798.9 m, with relative relief 1,711.9 m. The topography of the DJK area is overall higher to the northwest, and lower in the southeast, with steep terrain in the north and gentle slopes in the south, and alternating basins and canyons along the Han River (Bao, 2013). Situated in the transition zone to humid and warm climate within northern subtropical climate belt, DJK has a semi-humid continental climate, with four distinct seasons and precipitation is abundant in the wet season. Primary soil types in the DJK area include mountain yellow-brown earth soil, cinnamon soil, mountain brown soil, purple soil etc. The main forest and vegetation types are coniferous forests, broad-leaved forest, bamboo forest, shrub, and shrub meadow (Liao, 2011).

As the water source of the Central Route of the South-North Water Transfer Project, DJK has a storage capacity of 17.64 billion m^3, with averaged incoming runoff 39.35 billion

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1704_95619575

m^3. Incoming runoff occurs mainly during the wet season (July – October) with an estimation of more than 60% of the annual total in this season (Li et al., 2008, 2009; Yang et al., 2012). The catchment area of the reservoir area is formed by the convergence of the Han River and the Dan River. Its main tributaries include the Qianyou River, Jinqian River, Si River etc. The largest tributaries of the Dan River include the Qi River and Laoguan River. Hydraulic constructions built on the upstream of the DJK, many large reservoirs along the Han River, as shown in *Figure. 1*.

Figure 1. *The geography in the upstream of DJK and the distribution of hydrometeorological stations*

The Central Route of the South-North Water Transfer Project provides water supplies to more than 20 medium and large cities in Henan province, and Hebei province, including Tianjin and Beijing. Annual water transfer is estimated to reach ~9.5 billion m^3 by the end of the first phase project, and will reach ~13.0 billion m^3 in the medium and long term. The Central Route of the South_North Water Transfer Project is expected to significantly relieve the water shortage crisis in many regions of northern China (Chen et al., 2015).

Data sources

The climate indices dataset from CMA Climate Center (http://cmdp.ncc-cma.net/cn/monitoring.htm) includes 130 indices for atmosphere and ocean circulations. Correlation analysis (Keane and Adrian, 1993) is conducted between the annual mean runoff at DJK and individual climate indices at the previous year. In addition to these climate indices, accumulated precipitation at the previous year is also considered in our model as a predictor, because precipitation is one of the primary factors contributing to runoff.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1704_95619575

Gaussian Mixed Model (GMM)

GMM was developed based on Hidden Markov Model (HMM) (Eddy, 1996), and it belongs to a broad class of the unsupervised clustering methods (Reynolds, 2009). In essence, GMM is a multidimensional probability distribution function. Gaussian distributions are linearly weighted to characterize statistical distributions of the samples fully. Characteristic parameters spanning the space determine the model parameters (Mcnicholas and Murphy, 2008). In this paper, the actual runoff value of DJK from 1981 to 2016 was selected for clustering steps for clustering analysis in GMM may be summarized as follows:

(1) Suppose samples follow k mixed Gaussian distribution. Initialize μ_j , $\sum_j$, $j \in \{1,\cdots,k\}$ for each GMM.

(2) For each sample x^i , where $i \in \{1,\cdots,m\}$, compute the probability w_j^i that x^i follows GMM as following:

$$w_j^i = p\left(x^i \mid z^i = j\right) = \frac{1}{(2\pi)^{\frac{n}{2}} \left|\sum\right|^{\frac{1}{2}}} e^{\frac{1}{2}(x^i-\mu_j)\Sigma^{-1}(x^i-\mu_j)^T} \quad \text{(Eq.1)}$$

where z^i is the category of x^i .

(3) Update parameters of each Gaussian distribution:

$$\mu_j = \frac{\sum_{i=1}^m w_j^i x^i}{\sum_{i=1}^m w_j^i} \quad \text{(Eq.2)}$$

$$\sum_j = \frac{\sum_{i=1}^m w_j^i (x^i-\mu_j)(x^i-\mu_j)^T}{\sum_{i=1}^m w_j^i} \quad \text{(Eq.3)}$$

(4) Iterate steps (2) and (3) until the Gaussian parameters μ_j and $\sum_j$ converge.

(5) With the known Gaussian parameters μ_j and $\sum_j$ from the above step, iterate throughout all the samples and classify the samples according to the maximum probability.

Support Vector Classification (SVC)

The Support Vector Machines (SVM) is a type of the supervised classification methods developed from convex optimization (Vapnik, 1999; Hsu, 2010). SVC has been applied to classification and regression prediction problems. For classification, SVC can be grouped in two types: linear SVC and nonlinear SVC (Brereton and Lloyd, 2010). Linear SVC method solves the following optimization problem to identify the optimal classification interface:

$$\max_a L = \sum_{i=1}^N a_i - \frac{1}{2}\sum_{i,j=1}^N a_i y_i a_j y_j (x_i \cdot x_j) \quad \text{(Eq.4)}$$

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1704_95619575

where the objective function is subject to the following constrains:

$$\sum_{i=1}^{N} a_i y_i = 0, a_i \in [0, C], i = 1, \cdots, N \quad \text{(Eq.5)}$$

where a and a_i are Laplacian multipliers, c is the penalty factor.

Nonlinear SVC makes use of kernel functions and the optimal problem can be expressed as:

$$\max_{a} L = \sum_{i=1}^{N} a_i - \frac{1}{2} \sum_{i,j=1}^{N} a_i y_i a_j y_j K(x_i \cdot x_j) \quad \text{(Eq.6)}$$

The classifiers in linear and nonlinear SVC may be expressed as:

$$f(x) = sign\left(\sum_{i=1}^{N} a_i^* y_i (x \cdot x_i) + b^* \right) \quad \text{(Eq.7)}$$

$$f(x) = sign\left(\sum_{i=1}^{N} a_i^* y_i K(x \cdot x_i) + b^* \right) \quad \text{(Eq.8)}$$

where a_i^* is from the coupled optimal solution, and b^* is the offset coefficient. Commonly used kernel functions include: linear kernels, $K(x, x') = x\ x'$; polynomial kernels, $K(x, x') = [(x\ x') + 1]^d$; Gaussian radial basis function (RBF) kernel, $K(x, x') = \exp(-\left\| x - x' \right\|^2 / 2\sigma^2)$; where x, x' are vectors, and d is the degree of the polynomials, and σ is the bandwidth of the Gaussian kernel.

Results and Discussion

Figure 2 shows the annual mean incoming runoff at DJK from 1981 to 2016. It is evident that the annual runoff varies over a wide range. Therefore, it is difficult to accurately predict future values based on qualitative analysis of physical factors. This study develops a new model based on runoff classification and forecast domain to categorize annual runoff at DJK in the past 36 years, conduct forecast experiments, and perform validation of this forecast model.

Classification of runoff

Several methods have been used to classify runoff categories in the past: aggregate standard deviation, mean deviation, and percent deviation. Here we use the percent deviation method that calculated the distance percentage between the annual runoff value and the average of the 36 years (1981-2016) runoff to classify the DJK runoff data into two categories. If the percent deviation is negative, the year of runoff category is designated as 1; if it is positive, the category is designated as 2. *Table 1* lists the classification of runoff based on percent deviation into two categories derived from this method.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1704_95619575

Similarly, we further classify this DJK runoff data into 5 categories: Category 1 for percent deviation less than –20% , the percent deviation falls between –20% and –10%is Category 2 , Category 3 for percent deviation between 10% and –10%,. The percent deviation between 10% and 20% is Category 4 and greater than 20% is Category 5. Result from this percent deviation classification of the 36-year data is listed in *Table 2*.

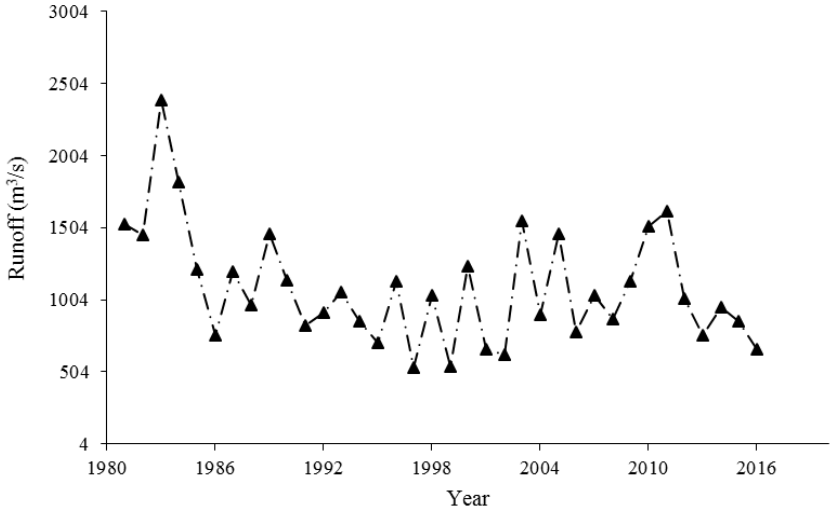

Figure 2. *Annual average runoff trend of DJK in 1981-2016*

Table 1. *Classification of runoff into two domains*

Year	Distance to average(%)	Domain	Years	Distance to average(%)	Domain
1981	40.51	2	1999	-50.16	1
1982	33.98	2	2000	13.56	2
1983	119.95	2	2001	-39.40	1
1984	67.72	2	2002	-42.53	1
1985	11.36	2	2003	42.89	2
1986	-30.57	1	2004	-17.15	1
1987	10.25	2	2005	34.16	2
1988	-10.80	1	2006	-28.09	1
1989	34.62	2	2007	-4.92	1
1990	4.83	2	2008	-20.18	1
1991	-24.32	1	2009	4.09	2
1992	-16.14	1	2010	39.40	2
1993	-2.99	1	2011	48.69	2
1994	-21.66	1	2012	-7.13	1
1995	-34.99	1	2013	-30.39	1
1996	4.18	2	2014	-12.74	1
1997	-50.80	1	2015	-21.10	1
1998	-4.75	1	2016	-39.40	1

Classification of forecast domain

Considering the distribution characteristics of the runoff sequence, the intrinsic distribution rule of the runoff sequence can be more effectively described, and the divided

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1704_95619575

runoff interval is more reasonable. Based on this assumption, we proposed a new method called forecast domain. With this method, the GMM is used to cluster the annual mean runoff. Gaussian probability distribution is then linearly weighted to fit the full statistical distribution, which reduces sampling noise and helps achieve better classification. The number of the iteration steps is set as 100, and the number of the clusters is 2. Forecast domain was constructed based on the results of the above cluster analysis. *Table 3* lists the results of the classification of forecast domain derived from this method.

Table 2. *Classification of runoff into five domains*

Year	Distance to average(%)	Domain	Year	Distance to average(%)	Domain
1981	40.51	5	1999	-50.16	1
1982	33.98	5	2000	13.56	4
1983	119.95	5	2001	-39.40	1
1984	67.72	5	2002	-42.53	1
1985	11.36	4	2003	42.89	5
1986	-30.57	1	2004	-17.15	2
1987	10.25	4	2005	34.16	5
1988	-10.80	2	2006	-28.09	1
1989	34.62	5	2007	-4.92	3
1990	4.83	3	2008	-20.18	1
1991	-24.32	1	2009	4.09	3
1992	-16.14	2	2010	39.40	5
1993	-2.99	3	2011	48.69	5
1994	-21.66	1	2012	-7.13	3
1995	-34.99	1	2013	-30.39	1
1996	4.18	3	2014	-12.74	2
1997	-50.80	1	2015	-21.10	1
1998	-4.74	3	2016	-39.40	1

Table 3. *Classification of forecast domain into two categories*

Year	Forecast domain	Year	Forecast domain	Year	Forecast domain
1981	2	1993	1	2005	2
1982	2	1994	1	2006	1
1983	2	1995	1	2007	1
1984	2	1996	1	2008	1
1985	1	1997	1	2009	1
1986	1	1998	1	2010	2
1987	1	1999	1	2011	2
1988	1	2000	2	2012	1
1989	2	2001	1	2013	1
1990	1	2002	1	2014	1
1991	1	2003	2	2015	1
1992	1	2004	1	2016	1

Similarly, GMM clustering of DJK annual runoff classifies the forecast domain into five categories. The number of the iteration steps is 100, and the number of the clusters is 5. *Table 4* lists the five forecast domains derived from this method.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1704_95619575

Table 4. *Classification of forecast domain into five categories*

Years	Forecast domain	Year	Forecast domain	Year	Forecast domain
1981	3	1993	2	2005	3
1982	3	1994	1	2006	1
1983	5	1995	1	2007	2
1984	4	1996	2	2008	1
1985	2	1997	1	2009	2
1986	1	1998	2	2010	3
1987	2	1999	1	2011	3
1988	2	2000	2	2012	2
1989	3	2001	1	2013	1
1990	2	2002	1	2014	2
1991	1	2003	3	2015	1
1992	2	2004	1	2016	1

Result Analysis

The CMA climate indices database is preprocessed in the following two steps. First, for a small number of missing values in this dataset, linear interpolation is used to fill these missing values. Second, a linear correlation is computed between the annual mean runoff at DJK and the climate indices at the previous year. Twenty climate indices that have the highest correlation with the DJK annual runoff are selected based on this lead correlation analysis. We use the 25-year data (1981-2005) for model training, and the DJK runoff data during last 11 years (2006-2016) are used for model validation. DJK has an Asian subtropical monsoon climate. Its precipitation mainly comes from two meteorological moisture sources: warm and humid moisture transport from the southeast and southwest (Guo and Jin, 1997). In addition, sea surface temperature at the Pacific Ocean and the Indian Ocean also play essential roles in the eastern Asia climate. Considering these factors, the following prediction factors are selected in *Table 5*: sea surface temperature anomalies at the NINO W region (September of the previous year), latitudinal position index of polar vortex center in Northern Hemisphere (May of the previous year), area index of warm pool in Western Pacific (July of the previous year), number of cold air (November of the previous year), location the subtropical high in South China Sea (December of the previous year), intensity of eastern Asian trough in June, and accumulated precipitation of DJK in the previous year. These variables are entered intothe SVC model. Prediction is then conducted for the conventional classification categories derived from the percent deviation method and forecast domain derived from the GMM classification method.

Table 5. *Correlation coefficients between the predictors and annual mean runoff*

No.	Factors	Correlation coefficient
1	Sea surface temperature departure index of NINO W district in September of the previous year	0.46
2	The latitude index of the polar vortex in the Northern Hemisphere in May of the previous year	0.46
3	Western Pacific warm pool area index in July of the previous year	0.43
4	The number of cold air in November of the previous year	0.42
5	The position index of the South China Sea subtropical high ridge in December of the previous year	0.42
6	East Asia trough intensity index in June of the previous year	0.40

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1704_95619575

Forecast and validation of runoff categories classification

To eliminate untended influence of the dimensions of various indices, the Min-Max Normalization method is utilized to normalize each index. *Figure 3* shows the forecast validation of the two classification categories from SVC. *Figure 4* shows the SVC forecast results of five classification categories. If the predicted results of a certain sample coincide with the real results, it indicates that the prediction is correct and the opposite is wrong. It is evident that ifthe classification of runoff into two categories, only the year 2009 and the year 2011 are predicted to be wrong, and the accuracy is 82%. However, if the classification of runoff is classified into five categories, the accuracy dropped to 45%.

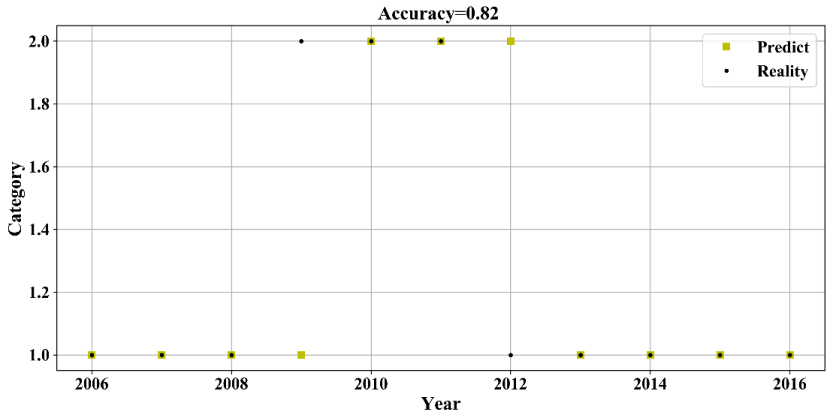

Figure 3. *SVC forecast based on two classification categories of runoff*

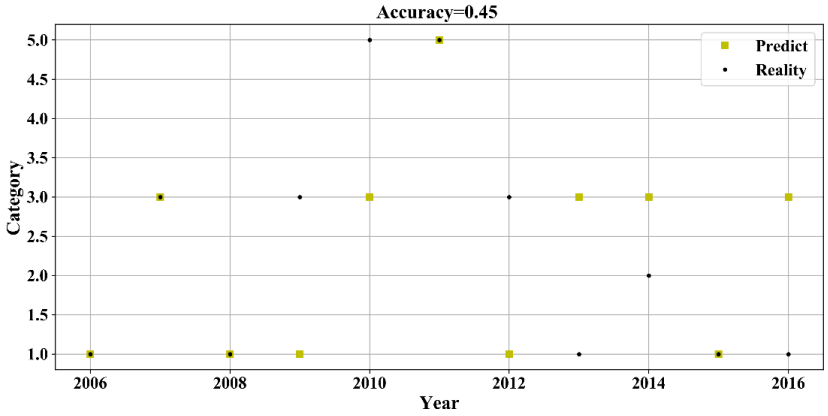

Figure 4. *SVC forecast based on five classification categories of runoff*

Forecast and validation of forecast domain

The similar forecast experiments are conducted. The forecast results based on SVC model by two and five classification categories of forecast domain are shown in *Figure 5* and *Figure 6*. The validation set represents the DJK runoff data during last 11 years (2006-2016) and the category means the classification of forecast domain. Compared with last section, when the classification of forecast domain into two categories, SVC forecast based on forecast domain performed better, and the accuracy

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1704_95619575

up to 91%. However, when the classification of forecast domain into five categories, the accuracy also declined. We can conclude that whether it is two categories or five categories, the accuracy based on forecast domain is higher than the original classification.

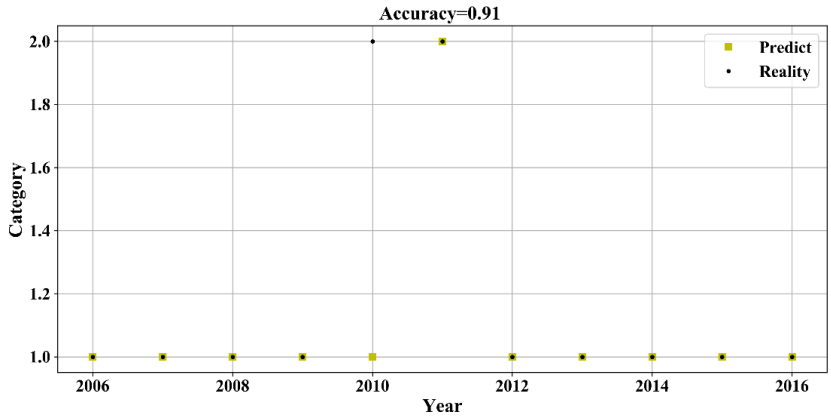

Figure 5. *SVC forecast based on two classification categories of forecast domain*

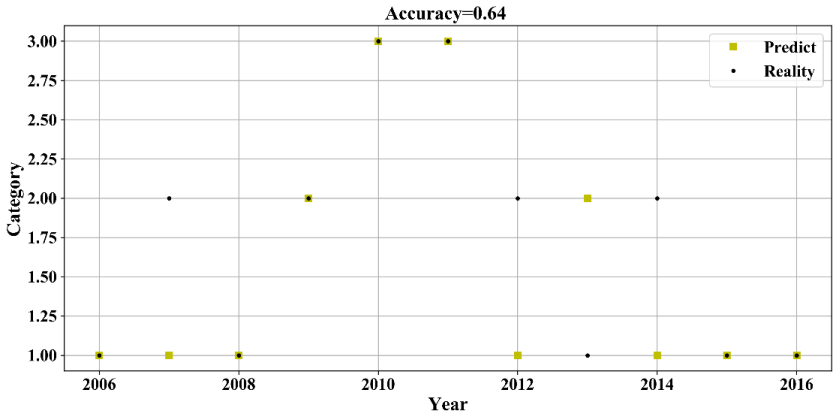

Figure 6. *SVC forecast based on five classification categories of forecast domain*

Table 6 summarizes the accuracy of SVC forecast based on theclassification methods by conventional percent deviation and forecast domain.

Table 6. *Comparisons of the SVC forecast by conventional percent deviation and forecast domain*

Accuracy	Two runoff categories	Two forecast domains	Five runoff categories	Five forecast domains
SVC	82%	91%	45%	64%

Reliability and performance of our prediction model are assessed with three forecast skill metrics: precision, recall rate (Buckland and Gey, 1994) and the F1 measure (Lipton et al., 2014). Precision indicates among the predicted results how many positive predictions are true positive. Precision is defined as:

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1704_95619575

$$Precision = \frac{number\ of\ true\ positives}{number\ of\ positives} \quad \text{(Eq.9)}$$

Recall measures among the relevant samples how many are correctly predicted, defined as:

$$Recall = \frac{correctly\ predicted\ number\ of\ true\ positives}{total\ number\ of\ relevant\ levels} \quad \text{(Eq.10)}$$

F1 is defined as the harmonic mean of precision and recall:

$$F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \quad \text{(Eq.11)}$$

These three skill metrics from these prediction experiments for the conventional classification categories and forecast domain are summarized in *Figure 7*.

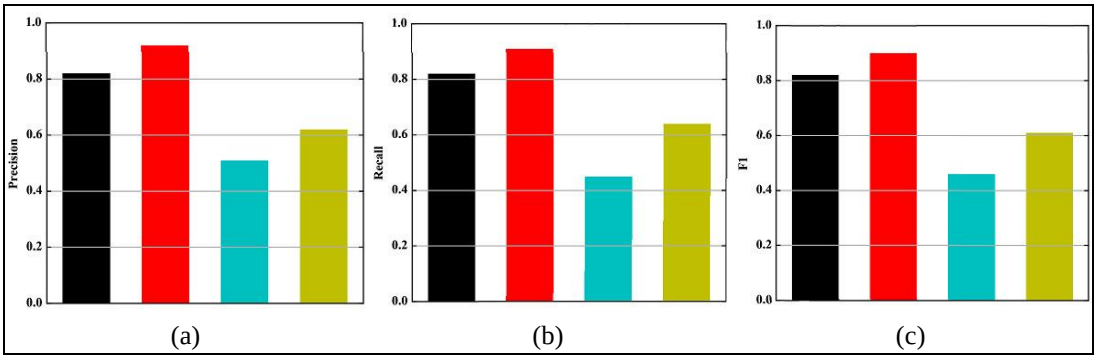

Figure 7. *Precision, recall, and F1 measures from the SVC forecast based on conventional percent deviation and forecast domain*

According to *Figure 3* and *Figure 4*, SVC prediction performs well for the two runoff categories derived from the percent deviation method, and the precision is 82%. However, the precision decreases to 45% by adopting the five runoff categories. In contrast, precision reaches 91% by using the two classification categories of forecast domain derived from the GMM clustering, and it remains 65% based on five classification categories of forecast domain.It is suggested that forecast domain method is more flexible and reliable. *Figure 7* further shows that regardless of the number of runoff categories, for all three metrics, SVC prediction based on forecast domain perform better than that based on conventional classification categories. Therefore, we conclude that the GMM clustering derived forecast domain is better suitable for the prediction of annual runoff at the DJK.

Model comparison

In order to better reflect the feasibility of this model, chooses Naive Bayes to compare with. The result is shown below.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1704_95619575

It can be seen from *Figure 8* and *Figure 9* that when two runoff categories performed, Naïve Bayes has a poor recall rate and is unstable compared with the SVC; when performing the five runoff categories, the accuracy of the two methods both are low and the effect is average. In summary, SVC has higher accuracy and stability.

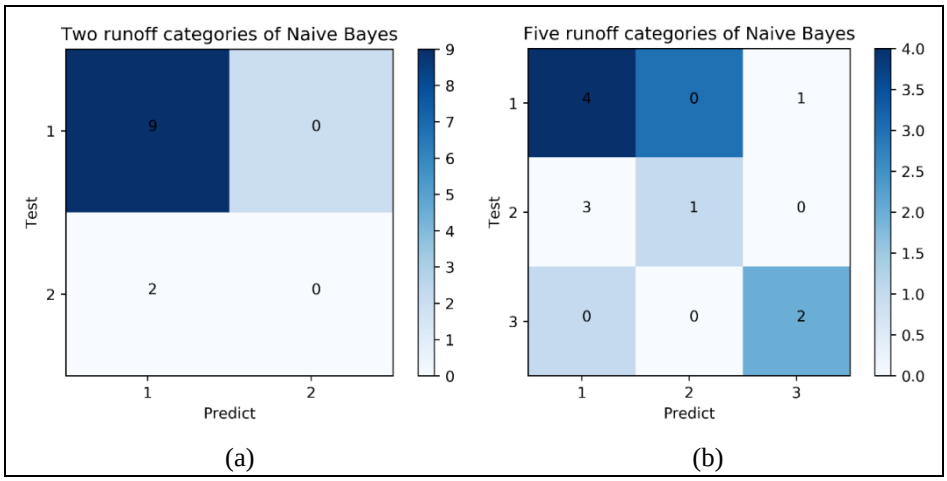

Figure 8. *Naive Bayes evaluation results*

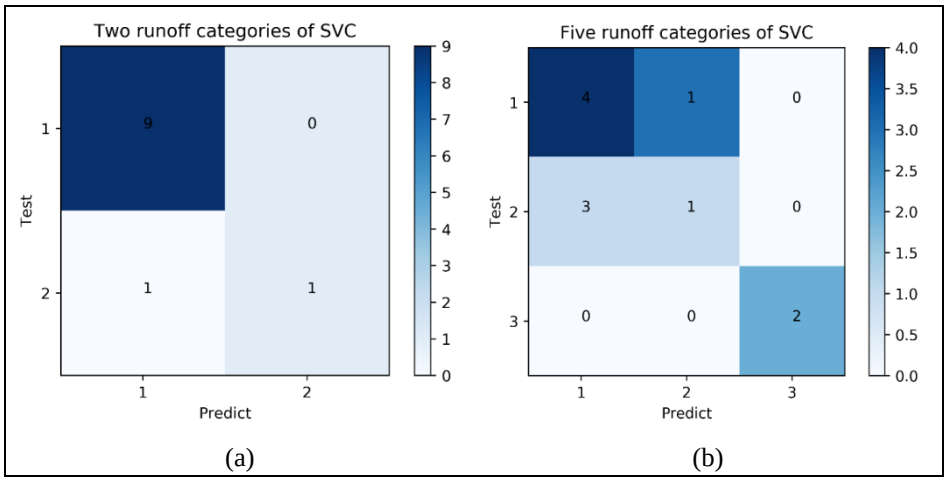

Figure 9. *SVC evaluation results*

Conclusion

The current study focuses on the DJK, the main water source of the Central Route of the South-North Water Transfer Project. We have developed a prediction model for annual mean runoff based on forecast domain instead of conventional classification method of runoff. Gaussian Mixed Model is utilized to cluster the 36-year annual runoff at DJK. Forecast domain was further derived following the GMM clustering and fed to the Support Vector Classification model. Forecast experiments based on forecast domain and conventional classification categories are conducted and compared. The results can be summarized as follows:

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1704_95619575

(1) Extending the prediction of single value to prediction of multiple forecast domain help better characterize and quantify the variability of runoff. This development bears important practical values to the water resource management and utilization.

(2) Application of sequence clustering analysis to classification takes consideration of the full distribution of runoff, which in turn helps better characterize the statistical distribution of runoff, and further justify objective classification of the runoff categories.

(3) Precipitation is one of the primary factors contributing to runoff. As a result, accumulated precipitation at the previous year should be included as a predictor.

Our results highlight that it is feasible to predict annual runoff based on forecast domain. This results may pave the way for the operational forecast of the annual runoff at the DJK. Our future research endeavors will be devoted to further improving the model in the following aspects:

(1) Using different feature selection method to characterize and understand potential nonlinear relationships among the predictors to achieve further forecast improvement.

(2) For the forecast domain developed in this study, we will apply the k-nearest neighbor's algorithm to predict future annual incoming runoff within a time range of interests at the DJK.

Acknowledgements. This paper was supported by National Key Research and Development Project (2016YFC0402201); National Science Foundation for Young Scientists of China (Grant No.51709271); Young Elite Scientists Sponsorship Program by CAST (2017QNRC001).

REFERENCES

[1] Bao, H. (2013): Analysis of the impact of the middle route project of south-to-north water diversion on the biodiversity of Danjiangkou reservoir area. – Northeast Forestry University, Harbin.

[2] Brereton, R. G., Lloyd, G. R. (2010): Support vector machines for classification and regression. – Analyst 135(2): 230-67. DOI:10.1039/b918972f.

[3] Buckland, M., Gey, F. (1994): The relationship between Recall and Precision. – Journal of the Association for Information Science & Technology 45(1): 12-19.

[4] Chen, P., Li, L., Zhang, H. B. (2015): Spatio-Temporal Variations and Source Apportionment of Water Pollution in Danjiangkou Reservoir Basin, Central China. – Water 7(6): 2591-2611.

[5] Coulibaly, P., Anctil, F., Rasmussen, P., Bobée, B. (2015): A recurrent neural networks approach using indices of low-frequency climatic variability to forecast regional annual runoff. – Hydrological Processes 14(15): 2755-2777.

[6] Eddy, S. R. (1996): Hidden Markov models. – Curr. Opin. Struct. Biol. 6: 361-365.

[7] Fan, Z. (1999): Medium and long-term hydrological forecast. – Hohai University Press, Nanjing.

[8] Guo, H. J., Jin, R. L. (1997): Status Quo and Change Trend of Water Resources in the Upper Reaches of Danjiangkou Reservoir. – Resources Science 24(1): 28-34.

[9] Hou, Z. Y., Lu, W. X., Song, W. B., Li, M. N., Chen, M. (2016): An annual rainfall forecast model based on ordered sample clustering for weighted Markov chains. – Systems Engineering Theory & Practice 36(4): 1066-1071.

[10] Hsu, C. W. (2010): A practical guide to support vector classification 67(5).

[11] Kasiviswanathan, K. S., Cibin, R., Sudheer, K. P., Chaubey, I. (2013): Constructing prediction interval for artificial neural network rainfall runoff models based on ensemble simulations. – J.Hydrol. 499(499): 275-288.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1704_95619575

[12] Keane, R. D., Adrian, R. J. (1993): Theory of cross-correlation analysis of PIV images. – Springer Netherlands 1-25.
[13] Li, S. Y., Gu, S., Liu, W. Z., Han, H. Y., Zhang, Q. F. (2008): Water quality in relation to land use and land cover in the upper Han River Basin, China. – Catena 75(2): 216-222.
[14] Li, S. Y., Cheng, X. L., Xu, Z. F., Han, H. Y., Zhang, Q. F. (2009): Spatial and temporal patterns of the water quality in the Danjiangkou Reservoir, China. – International Association of Scientific Hydrology Bulletin 54(1): 124-134.
[15] Li, S. Y., Zhang, Q. F. (2014): Partial pressure of CO2 and CO2 emission in a monsoon-driven hydroelectric reservoir (Danjiangkou Reservoir), China. – Ecol. Eng. 71(71): 401-414.
[16] Li, W., Zhou, J., Sun, H., Feng, K., Zhang, H., Tayyab, M. (2017): Impact of Distribution Type in Bayes Probability Flood Forecasting. – Water Resour. Manage. 31(3): 1-17.
[17] Liao, W. (2011): Research on Land Use Changes and Ecological Security Control in Danjiangkou Reservoir Area. – Central China Normal University, Wuhan.
[18] Lipton, Z., Elkan, C., Naryanaswamy, B. (2014): Optimal Thresholding of Classifiers to Maximize F1 Measure. – Springer Berlin Heidelberg 225-239 pp.
[19] Maslova, I., Ticlavilca, A. M., Mckee, M. (2016): Adjusting wavelet-based multiresolution analysis boundary conditions for long-term streamflow forecasting. – Hydrological Processes 30(1): 57-74.
[20] McNicholas, P. D., Murphy, T. B. (2008): Parsimonious Gaussian mixture models. – Statistics & Computing 18(3): 285-296.
[21] Quan, H., Srinivasan, D., Khosravi, A. (2014): Particle swarm optimization for construction of neural network-based prediction intervals. – Neurocomputing 127(6): 172-180.
[22] Reynolds, D. (2009): Gaussian Mixture Models. – Springer US, 93-105 pp.
[23] Tan, Q. F., Lei, X. H., Wang, X., Wang, H., Wen, X., Ji, Y., Kang, A. Q. (2018): An adaptive middle and long-term runoff forecast model using EEMD-ANN hybrid approach. – Journal of Hydrology 567: 767-780.
[24] Totz, S., Tziperman, E., Coumou, D., Pfeiffer, K., Cohen, J. (2017): Winter Precipitation Forecast in the European and Mediterranean Regions Using Cluster Analysis. – Geophys. Res. Lett. 44(24).
[25] Vapnik, V. N. (1999): An overview of statistical learning theory. – IEEE Trans. Neural Networks 10(5): 988-99.
[26] Yang, T. C., Yu, P. S., Chen, C. C. (2005): Long-term runoff forecasting by combining hydrological models and meteorological records. – Hydrological Processes 19(10): 1967-1981.
[27] Yang, Q., Xie, P., Shen, H., Xu, J., Wang, P., Zhang, B. (2012): A novel flushing strategy for diatom bloom prevention in the lower-middle Hanjiang River. – Water Res. 46(8): 2525-2534.
[28] Yang, T. T., Asanjan, A. A., Welles, E., Gao, X. G., Sorooshian, S., Liu, X. (2017): Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information. – Water Resources Research 53(4): 2786-2812.
[29] Ye, L., Zhou, J. Z., Zeng, X. F., Guo, J., Zhang, X. X. (2014): Multi-objective optimization for construction of prediction interval of hydrological models based on ensemble simulations. – Journal of Hydrology 519: 925-933.
[30] Zhao, T. T. G., Liu, P., Zhang, Y. Y., Ruan, C. Q. (2017): Relating anomaly correlation to lead time: Clustering analysis of CFSv2 forecasts of summer precipitation in China. – Journal of Geophysical Research Atmospheres 122(17): 9094-9106.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1704_95619575