Article ID: 1006-8775(2018) 02-0163-13

PREDICTION OF FLOOD SEASON PRECIPITATION IN SOUTHWEST CHINA BASED ON IMPROVED PSO–PLS

WANG Zhi-yi (王志毅)[1, 3], HU Bang-hui (胡邦辉)[1], YANG Xiu-qun (杨修群)[2], WANG Xue-zhong (王学忠)[1],
WANG Ju (王　举)[1], HUANG Hong (黄　泓)[1]
(1. College of Meteorology and Oceanography, PLA University of Science and Technology, Nanjing 210001 China;
2. School of Atmospheric Sciences, Nanjing University, Nanjing 210023 China;
3. Chongqing Meteorological Observatory,Chongqing 401147 China)

Abstract: In order to achieve the best predictive effect of the Partial Least Squares (PLS) regression model, Particle Swarm Optimization (PSO) algorithm is applied to automatically filter the optimal subset of a set of candidate factors of PLS regression model in this study. An improved version of the Particle Swarm Optimization-Partial Least Squares (PSO-PLS) regression model is applied to the station data of precipitation in Southwest China during flood season. Using the PSO-PLS regression method, the prediction of flood season precipitation in Southwest China has been studied. By introducing the precipitation period series of the mean generating function (MGF) extension as an alternative factor, the MGF improved PSO-PLS regression model was also built up to improve the prediction results. Randomly selected 10%, 20%, 30% of the modeling samples were used as a test trial; random cross validation was conducted on the MGF improved PSO-PLS regression model. The results show that the accuracy of PSO-PLS regression model and the MGF improved PSO-PLS regression model are better than that of the traditional PLS regression model. The training results of the three prediction models with regard to the regional and single station precipitation are considerable, whereas the forecast results indicate that the PSO-PLS regression method and the MGF improved PSO-PLS regression method are much better than the traditional PLS regression method. The MGF improved PSO-PLS regression model has the best forecast performance on precipitation anomaly during the flood season in the southwest of China among three models. The average precipitation (PS score) of 36 stations is 74.7. With the increase of the number of modeling samples, the PS score remained stable. This shows that the PSO algorithm is objective and stable. The MGF improved PSO-PLS regression prediction model is also showed to have good prediction stability and ability.

Key words: precipitation prediction; particle swarm optimization; partial least squares regression; flood season precipitation of Southwest China

CLC number: P426.6 **Document code:** A

doi: 10.16555/j.1006-8775.2018.02.005

1 INTRODUCTION

Southwest China locates in the transition zone from the Qinghai-Tibet Plateau to the plain south of the Yangtze River, with features of complex terrain and local synoptic factors and is also a typical climate change area (Hua et al.[1]). Because Southwest China belongs to the subtropical monsoon area, it is affected by the southwest monsoon significantly with distinct dry and wet season; relevant statistics show that more than 50% of precipitation in Southwest China originates from summer flood season. In recent years, affected by global warming, a series of drought and flood events with increased frequency and enhanced intensity occurred in Southwest China. In the summer of 2013, five strong storms hit Sichuan province persistently. Till 2014, six consecutive years of drought had been observed during winter and spring in Yunnan province. Along with the significant global warming since the end of the 1990s (Shen et al.[2]), Trenberth and Kharin et al. pointed out that an increase in the ground temperature intensifies the surface evaporation, which increases the moisture in the atmosphere, resulting in more possibility of precipitation [3-4]. On the other hand, enhanced evaporation of land surface makes the local drought more likely to occur and uneven distribution of rainfall also increases at the same time. After the analysis of the precipitation sensitivity of the spatial resolution in all the sub-regions over China, Zhong et al. found that among the climatic regions, it is the southwest region where the detection and prediction of climate variability is the most difficult [5]. The difficulty of precipitation prediction significantly increases due to the regional geography characteristics of the southwest region. The frequent occurrence of short term heavy rainfall events

Received 2017-05-12; **Revised** 2018-04-04; **Accepted** 2018-05-15

Foundation item: National Natural Science Foundation of China (41475070, 41375049, 41330420)

Biography: WANG Zhi-yi, M.S., primarily undertaking research on climate forecast and weather forecast.

Corresponding author: HU Bang-hui, e-mail: hubanghui@126.com

often leads to landslides, flash floods, and other geological disasters, which significantly threat human life and property.

The PLS regression method is an extension of the least square (LS) method, which was originally proposed by Wold et al. (1983) in the 1970s [6]. At the beginning of the 1980s, it was successfully used in the industrial field, after which its applications were quickly expanded to other areas. Compared with LS or other modeling methods, PLS regression model was simple, robust, efficient in its calculations, highly accurate, and did not need removal of any explanatory variables or sample points. Wold et al., Höskuldsson, and Geladi et al. point out that for the problem of constructing a regression model with multiple independent variables, when there are high correlation among the set of variables, modeling with partial least squares regression analysis is more effective than multiple regression, with more reliable forecast and stronger consistency overall[7-9]. Ertaç et al. use nonlinear time series and variable selection method to improve the PLS technology to predict the monthly mean air temperature in Istanbul[10]. They consider temperature, humidity, precipitation, and other elements as predictors and compare the improved PLS with ordinary least square, PLS, and artificial neural network prediction model and prove that the technology has higher prediction accuracy than that of the others.

At present, the forecast accuracy of rainfall in flood season based on climate numerical model is still under development (Kulkarni et al.[11]; Wang et al.[12]). Statistical model is an effective method for short term climate prediction. The initial multiple regression model is generally linear and the determination of its parameters depends on the LS method. However, there is an approximate linear relationship between multiple independent variables, which makes the regression equation unstable. Massy proposes the principal component regression method, which is based on biased estimates and overcomes issues with the estimation of the instability caused by the multicollinearity problem [13]. Hoerl proposes another geometric approach, the Ridge estimate(RE)[14]. Webster proposes the latent root regression, which is improved for multicollinearity[15].

The traditional regression technique often has a non-optimization problem, i.e. it cannot “rationally” filter out a number of factors from an alternative large number of independent variables to establish the so-called optimal regression equation. With the development of computational methods and advances in computers, Furnival and Sehatzoff propose an exhaustive "all possible" regression algorithm that greatly reduce the amount of computation required; as such, the problem of optimal regression could be solved more thoroughly [16-17]. However, the linear regression equation does not reflect the nonlinear relationship between the forecasting factor and the predictor. Therefore, the introduction of a nonlinear regression method is necessary.

PLS regression method solves the problem of collinearity (Wold et al. [18]; Abdi et al. [19]); it is also known as a new generation regression method. A prefect statistical model consists of the algorithm and the predictor selection. In the aspect of predictor selection, it is difficult to select the predictors automatically and artificial screening is required, which leads to the existence of subjective and personal dependence. In this study, a new prediction model is proposed, namely, PSO-PLS regression model. The method uses the PSO algorithm to automatically select the optimal factor combination automatically and objectively, and then the predictors are utilized and the precipitation is forecasted by PLS regression method. It inherits the excellence of PLS regression model, supplies the automatic selection of factors at the same time. In this article, it is applied to forecast the precipitation of flood season in Southwest China. The article is arranged as follows: data and evaluation criteria are introduced in section 1. In section 2, the basic theory and routine of establishment of PSO-PLS are described. The results and assessments of PSO-PLS are discussed in section 3. In section 4, PSO-PLS is compared with traditional PLS and the MGF-improved PSO-PLS. Random cross validation is also conducted to check the stability of the improved PSO-PLS model. Conclusions are drawn in section 5.

2 DATA AND EVALUATION CRITERIA

2.1 *Data*

The monthly average precipitation data of 160 stations during 1951—2014 in China is from the National Climate Center of CMA. Among them, 36 stations (Zhijiang, Liuzhou, Nanning, Beihai, Baise, Zunyi, Guiyang, Bijie, Xingren, Rongjiang, Enshi, Daxian, Youyang, Chongqing, Nanchong, Neijiang, Mianyang, Chengdu, Yibin, Ya’an, Kangding, Xichang, Huili, Lijiang, Dali, Baoshan, Kunming, Lincang, Meng, Jinghong, Ganzi, Deqin, Qamdo, Hanzhong, Ankang, and Yushu) are selected to represent entire Southwest China (97–110°E, 21–34°N) (Yan et al.[20]) (all stations distribution is shown in Fig.1). The regional average value of total precipitation from June to August of the 36 stations represents the entire region precipitation in flood season in Southwest China. Hereafter, precipitation during the flood season in Southwest China is shortened as PreFS.

The forecast factors are selected from the 126 indexes data offered by the National Climate Center. The indexes include the circulation index, SST index, zonal wind index, and teleconnection index. They extend from 1951 to 2014, with a summation of 64 years.

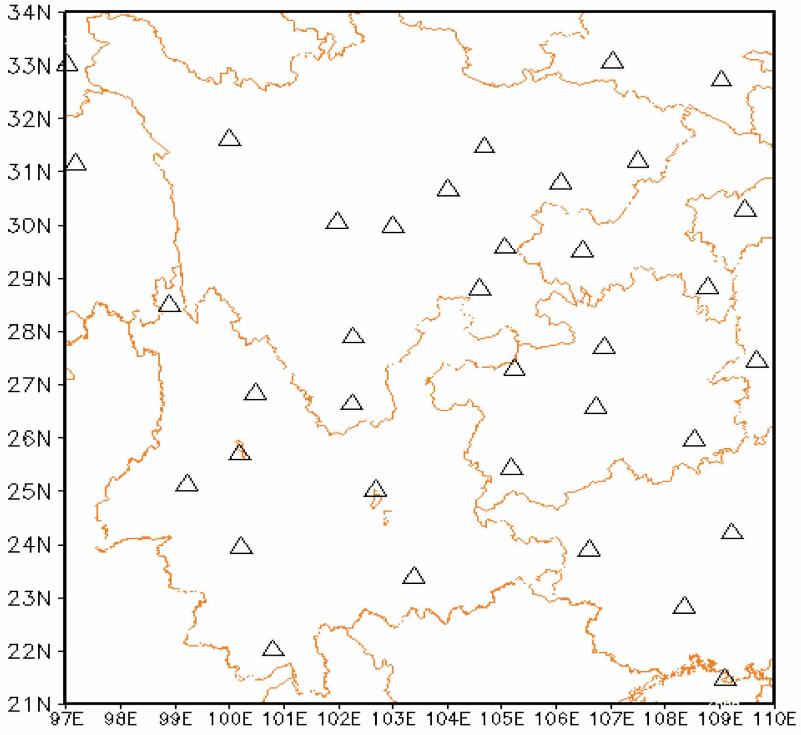


Figure 1. The distribution of 36 stations in Southwest China.

2.2 *Evaluation criteria*

In order to evaluate the prediction of stations and whole region PreFS values, root mean square error (RMSE), mean absolute error (MAE), anomaly correlation coefficient (ACC), and prediction skill score (PS) between the predicted and observed PreFS are calculated and compared.

RMSE is also known as standard deviation error. The root of the mean square deviation between the observed and the predicted values is shown as follows:

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}(X_{obs,i}-X_{for,i})^2}{n}} \qquad (1)$$

where n is the number of measuring times; $X_{obs,i}-X_{for,i}$ (i=1,2,3 ⋯ n) represents the deviation between the i-th pair of observed and predicted values.

MAE is the average of the absolute value of the deviation between the observed value and the predicted value. Because each sample of MAE is no less than zero (absolute value), MAE does not denote any positive or negative phase offset. MAE can better reflect the actual situation of the error between the prediction value and the observed value than average error whose positive deviation samples counteract the negative samples.

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}|X_{obs,i}-X_{for,i}| \qquad (2)$$

The symbols on the right hand side are similar to Eq. (1).

ACC is a criterion recommended by the World Meteorological Organization in 1996. It is the correlation coefficient between the predicted anomaly and the observed anomaly, which reflects the consistency of the predicted anomaly and the observed anomaly in space distribution. The expression is:

$$\mathrm{ACC}=\frac{\sum_{i=1}^{N}(\Delta R_{f,i}-\overline{\Delta R_f})(\Delta R_{o,i}-\overline{\Delta R_o})}{\sqrt{\sum_{i=1}^{N}(\Delta R_{f,i}-\overline{\Delta R_f})^2\sum_{i=1}^{N}(\Delta R_{o,i}-\overline{\Delta R_o})^2}} \qquad (3)$$

where N is the number of total stations; $\Delta R_{f,i}$ and $\Delta R_{o,i}$ are predicted and observed anomaly of the i-th station, respectively; $\overline{\Delta R_f}$ and $\overline{\Delta Ro}$ are the mean predicted and observed anomaly values of all stations, respectively.

PS score (Jin et al.[21]) is a widely used assessment in short term climate prediction in China to evaluate the forecast results of multiple stations. Its expression is

$$P_s=\frac{N_0+P_1\times N_1+P_2\times N_2}{N+P_1\times N_1+P_2\times N_2}\times 100 \qquad (4)$$

where P_s represents PS score, N is the total stations number, N_0 is the sum number of the normal level stations, which include the same or counter symbol between the forecast and the observation anomaly. N_1 is the number of stations with corrected prediction of the first level anomaly. N_2 is the number of stations with corrected prediction of the second level anomaly. P_i represents the weight coefficient, P_1=5 and P_2=2 represent the weighing parameters of successful prediction of the first and second level anomaly, respectively. Table 1 lists the trend forecast diction and classification standard of a five-level system of rainfall.

Table 1. Glossaries of trend forecasting and classifying criterion in a 5-degree prediction skill (PS) scheme.

Level	Diction (precipitation)	Percentage of precipitation anomaly/% (seasonal scale)
1	Extreme drought	$\Delta R \leqslant -50$
2	Moderate drought	$-50 < \Delta R \leqslant -20$
3	Normal	$-20 < \Delta R \leqslant 20$
4	Moderate flood	$20 \leqslant \Delta R < 50$
5	Extreme flood	$\Delta R \geqslant 50$

3 ESTABLISHMENT OF THE PREDICTION REGRESSION MODEL PSO-PLS

3.1 *Screening of predictive factors*

PSO algorithm, also named as particulate swarm algorithm, was proposed by Kennedy et al. based on the group behavior of birds and fish in 1995 [22]. The basic principles derive from the theory of artificial life and evolutionary computation. By imitating the flying and foraging behaviors of birds, it is expected to carry out the searching global optimization solution. The core idea is to use the birds collective collaborate to achieve the best gain of the group (Eberhart et al.[23]; Kennedy et al.[22]). In this study, this algorithm is used to choose the predictors at the initial stage.

At the moment of t+1, the speed and position update formula of PSO algorithm are as follows:

$$v_{id}^{t+1}=\omega\times v_{id}^{t}+c_1\times r_1\times(pBest_{id}^{t}-x_{id}^{t})+c_2\times r_2\times(gBest_{gd}^{t}-x_{id}^{t}) \quad (5)$$

$$x_{id}^{t+1}=x_{id}^{t}+v_{id}^{t+1} \quad (6)$$

In Eq. (5), v_{id}^{t} is the d-dimensional velocity of the particle i at iteration time t; c_1 and c_2 are accelerating coefficients (also known as learning factors), which adjust the maximum step size of the particle flight toward the individual best particle and the global best particle, respectively. If the acceleration coefficient is too small, the particles may be too far away behind to reach the target area. If it is very big, it may fly to the target area suddenly or fly over the target area (Eberhart at al.[24]). The suitable magnitudes of c_1 and c_2 can accelerate convergence and the solutions are not easy to fall into local optimum. Usually define c_1=c_2=2; r_1 and r_2 are random numbers in the range of [0, 1] (Kennedy et al.[25]).

The x_{id}^{t} is the current d-dimensional position of the particle i at iteration time t. $pBest_{id}$ is the position of the individual's d-dimensional extreme best points of the particle i. $gBest_{gd}$ is the position of the global d-dimensional extreme best point of the whole group. In order to prevent the particles from leaving the search space, each of the particles in the one-dimensional velocity v_d is limited within the range $[-v_{d\max}, +v_{d\max}]$. If $v_{d\max}$ is too large, particles will fly away from the best solution; too small particles will tend to fall into the local optimum (Eberhart et al. [24]). ω is the inertia weight, which is used to optimize the search for the solution. The related research shows that the larger inertia weight is beneficial to the global optimization, and the smaller inertia weight is beneficial to the local optimization. The iterative formula of inertia weight coefficient is

$$\omega_t=\omega_{\max}-t\times(\omega_{\max}-\omega_{\min})/t_{\max} \quad (7)$$

In Eq. (7), t represents the iteration number, $t_{\max}$ represents the maximum number of iterations. $\omega_{\max}$ and $\omega_{\min}$ are the maximum and minimum inertia weight coefficients, they are assigned to 0.9 and 0.4, respectively (Liu et al.[26]).

The PSO algorithm is also required to select the appropriate fitness function to evaluate the alternative particle's efficiency. Fitness function is defined as the reciprocal of quadratic sum of training error between normalized observed and predicted precipitation:

$$F=\frac{1}{\sum_{i=1}^{N}(R_{o,i}-R_{s,i})^2} \quad (8)$$

In Eq. (8), N is the station number and $R_{o,i}$ is the normalized value of the observed precipitation of stations. $R_{s,i}$ is the fitted normalized precipitation of stations.

Because the goal is to select factors, position vector is stored through binary coding. Each component of the position vector of the particle can only take the value of 0 or 1. The velocity vector of the particle represents the probability of being 1 at the next position. This class of PSO solving the discrete space problem is also known as binary particle swarm optimization (BPSO) (Kennedy et al.[27]; Ying et al.[28]). BPSO generally uses the Sigmoid function to handle . Due to the original Sigmoid function making BPSO strongly random, the lack of local detection ability makes it difficult to converge to the global optimal position. In this paper, the improved Sigmoid function is used to enhance the local detection ability of BPSO, which is beneficial to the convergence of the particle to the optimal position of the population (Liu[26]).

The improved Sigmoid function formula is (Liu[29]):

$$s(v_{id})=\begin{cases}1-\dfrac{2}{1+\exp(-v_{id})}, & v_{id}\leqslant 0\\[2ex] \dfrac{2}{1+\exp(-v_{id})}-1, & v_{id}>0\end{cases} \quad (9)$$

The position change formula for BPSO is (Liu[29]):

$$x_{id}^{t+1}=\begin{cases}1, & \mathrm{r}<\mathrm{sig}(v_{\mathrm{id}}^{\mathrm{t}})\\ 0, & \mathrm{r}\geqslant\mathrm{sig}(v_{\mathrm{id}}^{\mathrm{t}})\end{cases} \quad (10)$$

In Eq. (10), r is a random number between 0 and 1.

PSO algorithm is used to select the optimal subset consisting of 4–6 predictors from the primary circulation factors. The main process of screening is as follows:

(1) Candidate predictors for selection are chosen from the 126 indexes of the particular months of the present year (January and February) and the previous year (From March to December). Correlation coefficients are calculated, and the most relevant factors of the summer flood season precipitation of each station are selected based on the threshold correlation coefficient 0.3 and the number of factors is limited to no more than 15. These factors make up an alternative population.

(2) Initialization. The initial search point x_i^0 and speed v_i^0 are generated randomly within specified range. The $pBest$ coordinates of each particle are set at their present position. The corresponding individual extremum (i.e., the fitness value of the individual extreme points) is calculated. The global extremum (i.e., the fitness value of the global extreme points) is the best value among all the individuals. Then, the particle number of the best value is recorded, and the $gBest$ is set to the present position of the best particle.

(3) Assessment of each particle. The fitness value of each particle is calculated. If the fitness value is better than the current individual extremum, $pBest$ is renewed to the position of the corresponding particle, and the individual extremum is updated to the fitness value. If the individual extremum of all particles is better than the present global extremum, then $gBest$ is set to the position of the individual particle, and the number of the particle is recorded. The global extremum is updated similarly.

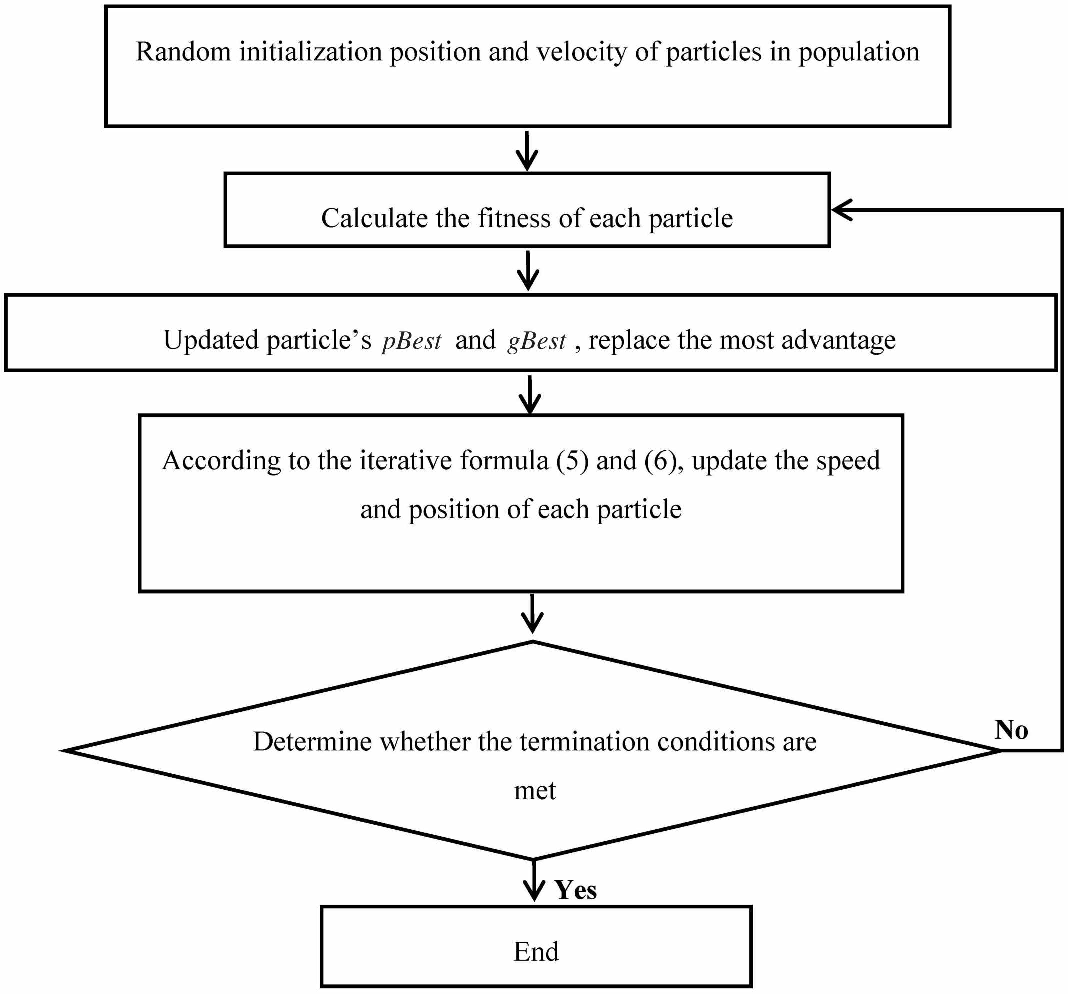

Random initialization position and velocity of particles in population
Calculate the fitness of each particle
Updated particle’s pBest and gBest , replace the most advantage
According to the iterative formula (5) and (6), update the speed and position of each particle
Determine whether the termination conditions are met
No
Yes
End

Figure 2. Particle swarm optimization (PSO) selecting factors flow chart.

(4) Particle update. According to the iterative formula (5) and (6), the velocity and position of each particle are iteratively updated.

(5) Termination check of iteration. If the current iteration number is up to the maximum number of preset times or the error of iteration less than the preset minimum threshold error, the iteration is stopped, then the optimal solution is output with the

optimal combination of predictors containing 4–6 members. If the termination conditions are not met, switch to step (2).

Using the PSO algorithm to select predictors can liberate humans from factor screening, and exclude those factors that have a good correlation with the predictor but are poor for modeling or forecasting. The algorithm requires little in computational resource and can filter out the best combination of factors automatically. Thus, it has good efficiency on factor screening.

3.2 *Establishment of the forecasting model*

Multicollinearity defects often appear in multi-factor prediction problems. The PLS regression model can solve it (Wold et al. [30]). By using PLS regression to obtain the final prediction equation coefficients, the prediction equation of PreFS at each station or of the entire region is established. The modeling steps of PLS are as follows.

First step: The optimal combination of predictors selected through PSO algorithm is substituted into the flood season precipitation training equation. Then, the precipitation samples are normalized, and the PLS regression method is used to extract the first component of the sample, and the cross validity test Q_1^2 is obtained (Wang [31]) (please refer to the reference section for the calculation of Q_1^2).

Second step: If $Q_1^2 \geqslant 0.0975$, it shows that the introduction of new principal components has a significant improvement in the predictive power of the model; in this case, the first step is repeated. Otherwise, the solution to the main component of the cycle process is over.

Third step: After determining the number of the principal components, the regression coefficients of the predictors are obtained, and the forecasting equation is determined.

Fourth step: Taking the PrePS (precipitation in the previous season) as the historical observation data and the predictors selected through the PSO algorithm, the prediction model is established. The fitness or predicted results of the PrePS are obtained.

4 RESULTS OF THE MODEL AND TEST

4.1 *Fitness value analysis of optimal subset*

Each optimal subset has commonality across the fitness values (from low to high) during the iteration process. For example, at Zunyi station, we analyzed the evolution of the characteristics of the fitness value of predictors optimal subset in the process of iterative using particle swarm algorithm. Fig.1 shows the curve fitting of fitness value of Zunyi station. As can be seen from Fig.1, after the iteration time reach 23, the fitness function value entered a state of convergence and the corresponding fitness value is 0.0394. The fitness function value increased to 0.0394 from 0.0287, i.e., an increase of 0.0107. Therefore, the predictor's optimal subset selected using the particle swarm algorithm has better fitting effect when compared with the general predictors subset.

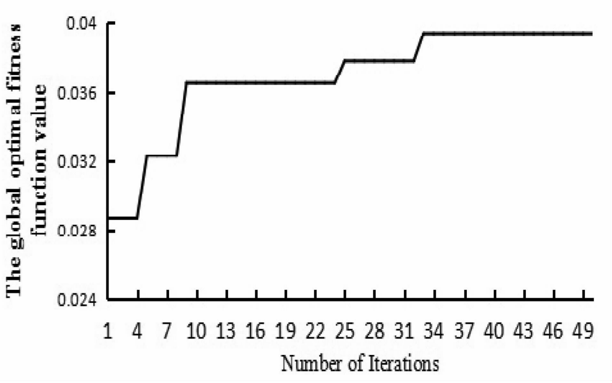


Figure 3. The simulated precipitation prediction fitness functions of PSO-PLS at Zunyi, the abscissa denotes PSO iteration times, and the *y*-axis denotes the global optimal fitness function.

4.2 *Prediction results of the model*

The PSO-PLS model is used to fit and predict the PreFS of 36 stations and the whole region. A total of 54-year section of dataset, which consists of PreFS from 1952 to 2005 and corresponding circulation indexes, are used. First, predictors are selected by PSO algorithm. Then, the whole region precipitation forecast model is established. And a 9-year (2006—2014) predictive experiment of the PreFS is carried out.

Figure 4 shows the training and forecast results of PSO-PLS model, together with the corresponding observation of PreFS. Table 2 shows the assessment statistics of the results. Modeling forecast results show that the forecast ACC is 0.31. This is acceptable even though the training result is much better comparatively. Moreover, RMSE, MAE, and ACC of the fitting results between 1952 and 2005 are 29.72, 24.21, and 0.79, respectively. The RMSE, MAE, and ACC of forecast results between 2006 and 2014 are 68.08, 51.67, and 0.31, respectively.

Table 2. The statistical assessments fitted/forecasted regional rainfall of flood season over southwestern China by PSO-PLS regression method.

	Fitted value	Forecast value
RMSE(mm)	29.72	63.08
MAE(mm)	24.21	50.67
ACC	0.79	0.31

In addition, the PSO-PLS models are established for 36 stations separately. The modeling data are from 1952 to 2008, and the forecast experiment temporal range is from 2009 to 2014. Forecast results show that the 36 stations mean value of test ACC is 0.32, which is similar to the entire region value. These results show that the forecast ability of PSO-PLS is acceptable and stable. The mean value of the PS score for forecasting the flood season precipitation in Southwest China during the flood season is 73.07, with the highest score of 87.5 in 2009 and the lowest of 62.5 in 2011 (as shown in Table 6).

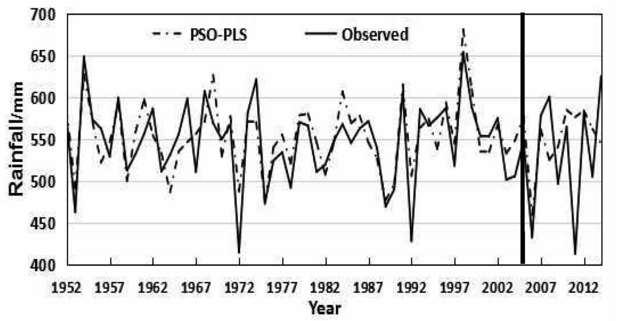


Figure 4. The yearly observed and fitted/forecasted regional rainfall during the flood season over southwestern China by PSO-PLS regression method (The fitness period is from 1951 to 2005, and the forecasting period ranges between 2006 and 2014. The periods are separated by a vertical line in the panel).

5 COMPARISON AND IMPROVEMENT OF THE MODEL

5.1 *Comparison of models*

In order to make a more intuitive description of the improvement of PSO-PLS, it is compared with the PLS model without any optimization. The PLS models are established for 36 stations separately and for the regional PreFS, then the corresponding training and prediction results are obtained. Model training period is 1952—2005, and the trial period is 2006—2014.

For the case of the whole region precipitation, the training and forecast results of PLS forecasting model are shown in Fig.5, together with the comparison observed PreFS. Table 3 shows the inspection results of prediction results by PLS model. Results show that the training ACC during 1952—2005 is 0.78, and the predictive ACC is 0.15. By comparing with the corresponding values in Table 2, only small differences are evident between the fitting results of the two models, whereas the forecast results of PSO-PLS regression model are much better than the traditional PLS regression model.

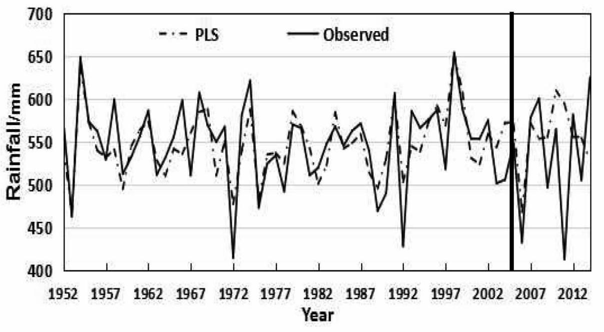


Figure 5. Same as Fig.4, but for the PLS regression method.

Table 3. Same as Table 2, but for PLS regression method.

	Fitted value	Forecast value
RMSE(mm)	29.96	74.05
MAE(mm)	23.52	56.89
ACC	0.78	0.15

The 36 stations mean value of forecast ACC based on the PLS regression model is 0.24 (the corresponding value of PSO-PLS is 0.32). The mean value of the PS score for PreFS is 71.1 (the corresponding value of PSO-PLS is 73.07), the highest score is 80.8 in 2014, and the lowest is 59.5 in 2011 (as shown in Table 6). The comparison of mean ACCs and PS scores indicates that PSO-PLS has advantage of PLS model.

A comparison between the modeling results of PSO-PLS method and PLS method is shown in Table 4. RMSE1 and MAE1 and RMSE2 and MAE2 represent the average values of the root mean square error and mean deviation of training and forecast cases, respectively. It shows that the fitting results of two methods are considerable. However, the forecast results show that PSO-PLS is superior to the PLS method in terms of RMSE, MAE, ACC, and PS score.

Table 4. Comparison between PSO-PLS and PLS on the statistical quantities of fitted/forecasted regional rainfall of flood season over southwestern China by PSO-PLS regression method.

	RMSE1	RMSE2	MAE1	MAE2	ACC	PS score
PLS	29.96	74.05	23.52	56.89	0.15	71.10
PSO-PLS	29.72	63.08	24.21	50.67	0.32	73.07

5.2 *Improvement of model*

As depicted above, the PSO-PLS is better than PLS method on predictive efficiency. However, it is deficient for extreme values such as the anomalous precipitation levels observed in 2011 and 2014; as shown in Fig.4, the difference between the predicted value and the observed value is still large. The possible reason is that the number of candidate predictors is insufficient to meet the requirement. In order to improve the predictive ability for extreme values, the precipitation time series of the certain period were extracted using the mean generation function (MGF). Each series of different periods from MGF were treated as candidate predictors to enrich the number of factors.

The MGF (Wei et al.[32]) is based on the definition of time series of the mean value generating function and its extension matrix. A set of periodic functions are constructed based on the original data sequence. Among the periodic functions, those representing the intrinsic period features of the precipitation are treated as candidates and participate in the selection of the PSO algorithm with other circulation predictors for PLS modeling.

The periodic factor is constructed through a MGF model (Wei[33]). The maximum period M = [N/3] ([ ] relates to rounding, N is sample number) is selected. From the previous modeling information, it can be found that N = 54, thus M=18. Therefore, 18 extension sequences of MGF are constructed. Their correlation with the PreFS time series is shown in Fig.6. The extension sequences of L = 14a (L represents period) has the highest correlation to the PreFS (r = 0.655) with significant level of α=0.01. The correlation coefficient between the extended sequences of L = 7a and the PreFS is 0.333, with significant level of α=0.05. The wavelet analysis of PreFS time series shows that the quasi periodic of 2a, 3–4a, and 5–7a on the interannual scale, and 14a on the decadal scale (figure omitted). Therefore, the extensions of L = 7a and L = 14a are used as alternative periodic factors for the forecasting model. This MGF improved PSO-PLS model is used to fit and forecast within the training period of 1951—2005 and the trial period of 2006—2014.

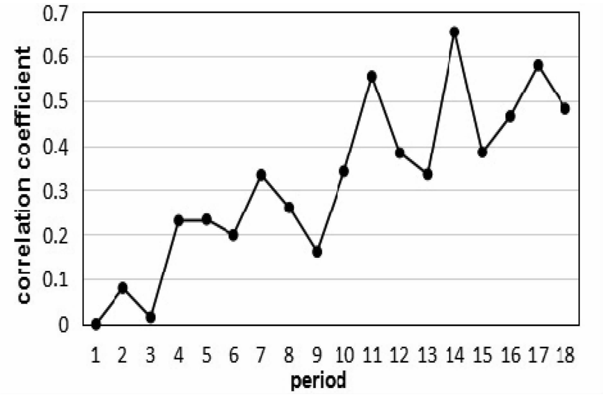


Figure 6. The correlation between a group of series of different periods produced by mean generated function (MGF) and their original observed regional rainfall in flood season over southwestern China.

PSO-PLS regression prediction model was improved through the addition of the MGF periodic predictors. The comparison of the training and forecast results of improved PSO-PLS forecasting model is shown in Fig.7. Table 5 shows the inspection results of prediction results of the improved PSO-PLS model. Fig. 6 shows that the difference between the training curve and the measured curve is larger than that of the first two models, indicating the poor training ability. However, the difference between the predicted curves and the observed curves of the precipitation anomalies during the forecast period is smaller than that of the prior two models. Modeling test results show that the training ACC during 1951—2005 is 0.69 and the forecast ACC is 0.44. Compared to the PSO-PLS method, the training effect of the MGF improved method is slightly worse but the prediction accuracy is improved obviously.

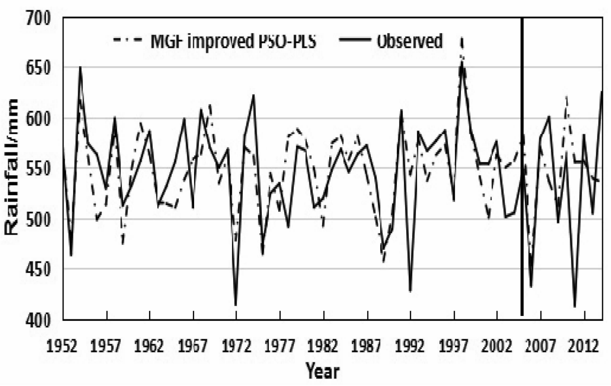


Figure 7. Same as Fig.4, but for the MGF improved PSO-PLS regression method.

The improved PSO-PLS model was established for the prediction of 36 stations separately. The 36 stations' mean test ACC is 0.33 and the mean PS score (Table 6) is 74.7, the highest score is 86.3 in 2012, and the lowest is 63.2 in 2010. The forecast score of anomalous precipitation in 2011 is 75, much higher than the other two methods' PS score. Obviously, the addition of the MGF extension sequences effectively improves the accuracy of the PSO-PLS model on prediction of the PreFs anomaly.

Table 5. Same as Table 2, but for MGF improved PSO-PLS regression method.

	Fitted value	Forecast value
RMSE(mm)	35.46	62.53
MAE(mm)	27.24	48.77
ACC	0.69	0.44

Table 6. PS score of three models of 36 stations precipitation test forecast results in Southwest China during flood season.

Year	2009	2010	2011	2012	2013	2014	Average Value
PLS	78.9	66	59.5	75	65.7	80.8	71.1
PSO-PLS	87.5	64	62.5	85	70.5	69	73.07
The MTF improved PSO-PLS	81.8	63.2	75.0	86.3	76.1	66	74.7

5.3 *Comparison of spatial distribution of three models*

In order to more intuitively compare the predicted results of the three models of flood season precipitation, Fig.8 shows the spatial anomaly rainfall percentage distribution of the predicted results for three models in 2012 and the observations of the spatial anomaly rainfall percentage distribution in 2012. As can be seen, the spatial characteristics of the precipitation in the southwest in 2012 mainly show greater precipitation in southeastern area, with the north and south-central areas showing less precipitation. The three models are basically able to reflect this spatial distribution, but there are differences in the magnitude of the precipitation. The forecast values of three models in the northwest and midwest regions are all less than the observations. The amounts of precipitation in the northeast area that were forecasted using the PLS model and PSO-PLS model are greater than the observed precipitation. The results of the MGF improved PSO-PLS model are closest to the observed values. As given by the three models, the amounts of precipitation in the southwestern area are greater than the observed precipitation, with the results of the MGF improved PSO-PLS model closest to the observed values.

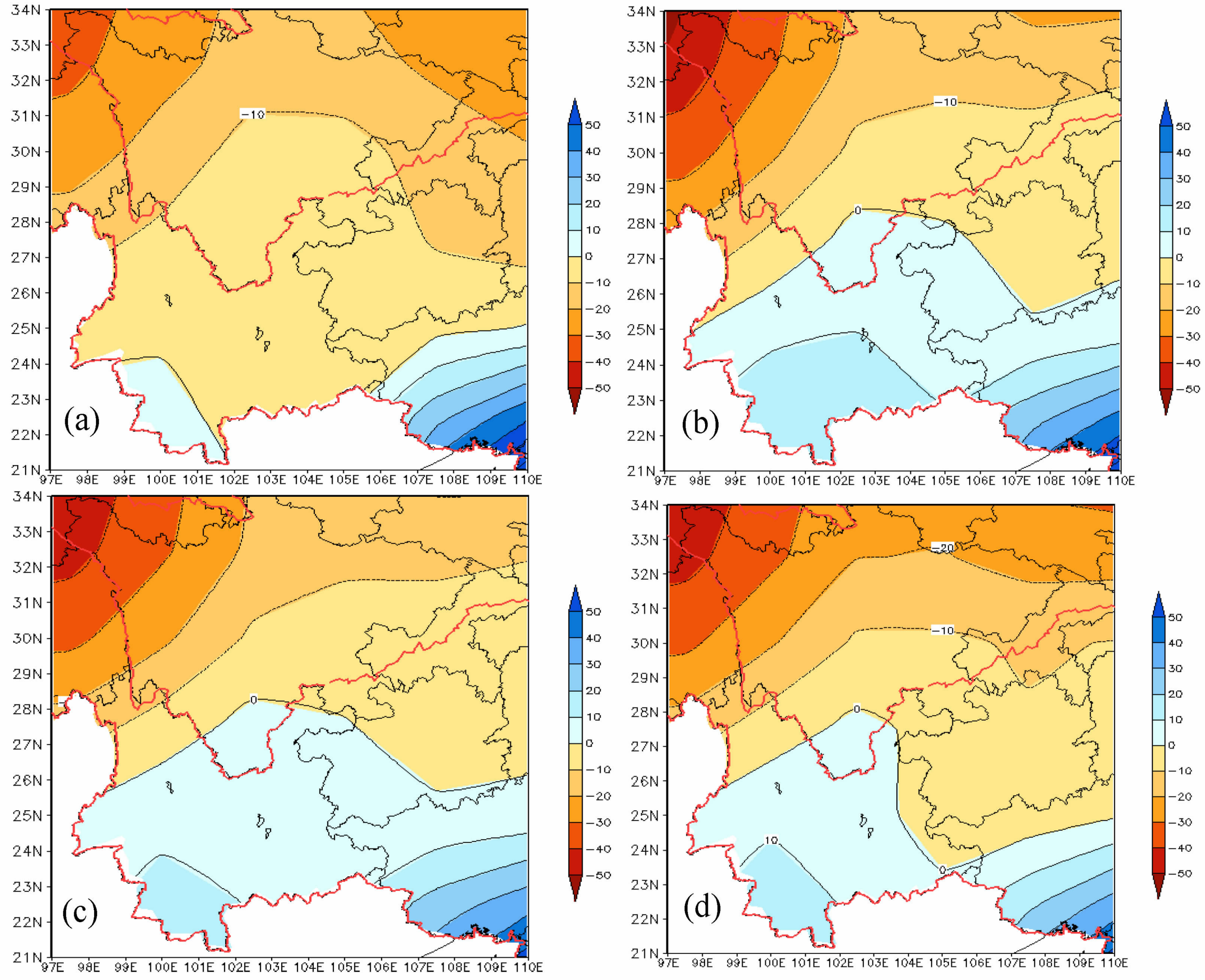


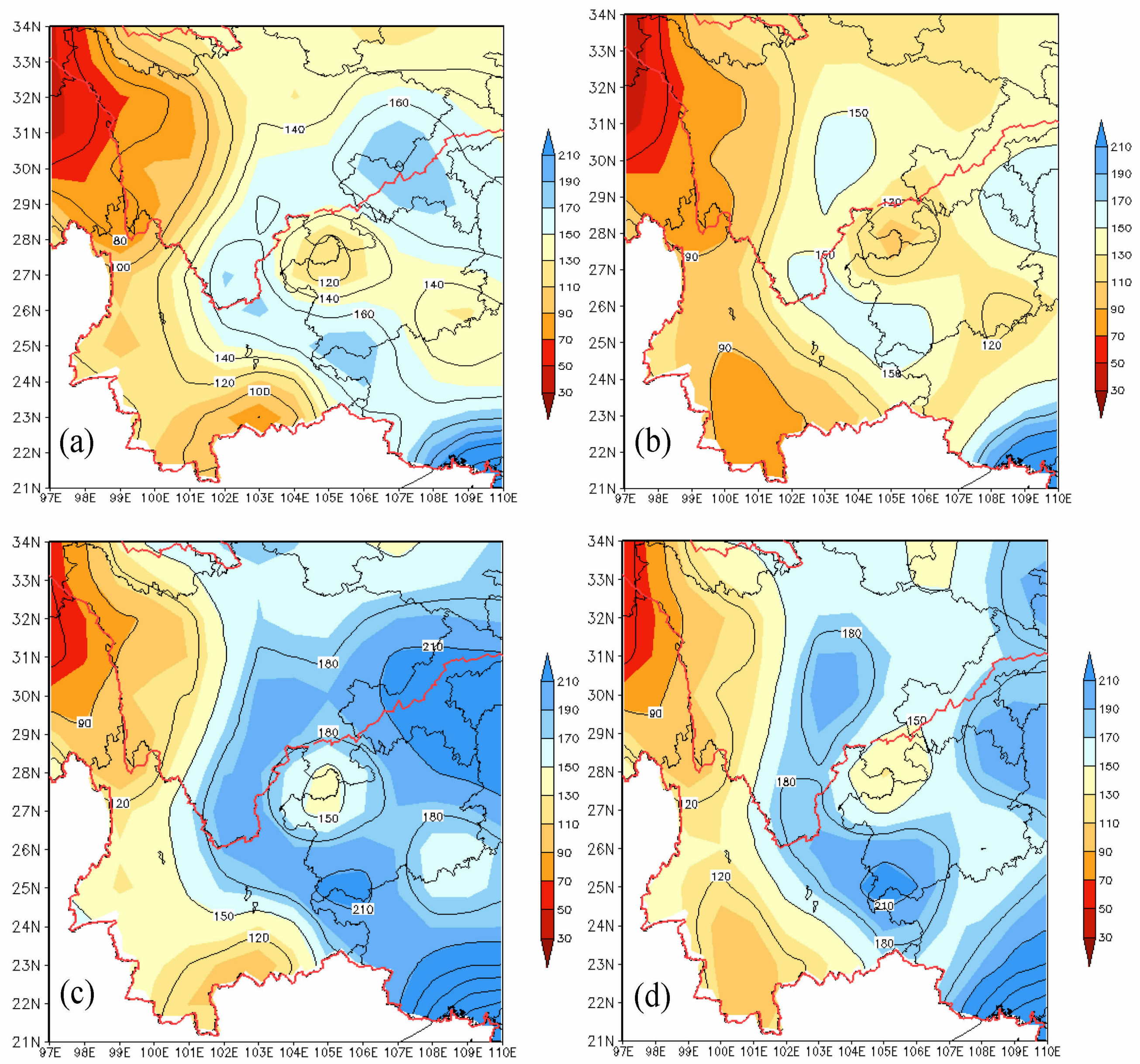
Figure 8. The spatial anomaly rainfall percentage distribution for three models and the observation of spatial anomaly rainfall percentage distribution in Southwest China. (a) observation (b) PLS model (c) PSO-PLS model (d) MGF improved PSO-PLS model.

In summary, it can be assumed that among the three models, the spatial distribution of the predicted results of the MGF improved PSO-PLS model is closest to the observed values. This validates the improved effect of the MGF improved PSO-PLS model on the predictive accuracy of precipitation in the PLS regression model.

In addition, Fig.9 shows the spatial distribution of the assessment scores of MAE, RMSE, ACC of the PLS model and the MGF improved PSO-PLS model. According to the comparison results, we can infer that the predictive ability of the MGF improved PSO-PLS model is generally stronger than that of the PLS model.

6 STABILITY OF THE MODEL

The previous analysis shows that the MGF improved PSO-PLS model is more effective in predicting PreFS. Using the random cross validation (Tong [34]; Xiong et al.[35]), the stability of the model is tested. 10%, 20%, and 30% samples are randomly selected from the 64 modeling samples of 36 stations, and the rest of samples are taken as the test samples. The sensitivity of the PSO algorithm to the predictor number of 13, 15, and 17 is also investigated. The average PS score of 36 stations from 20 times random cross validation of the forecast model are shown in Table 7.



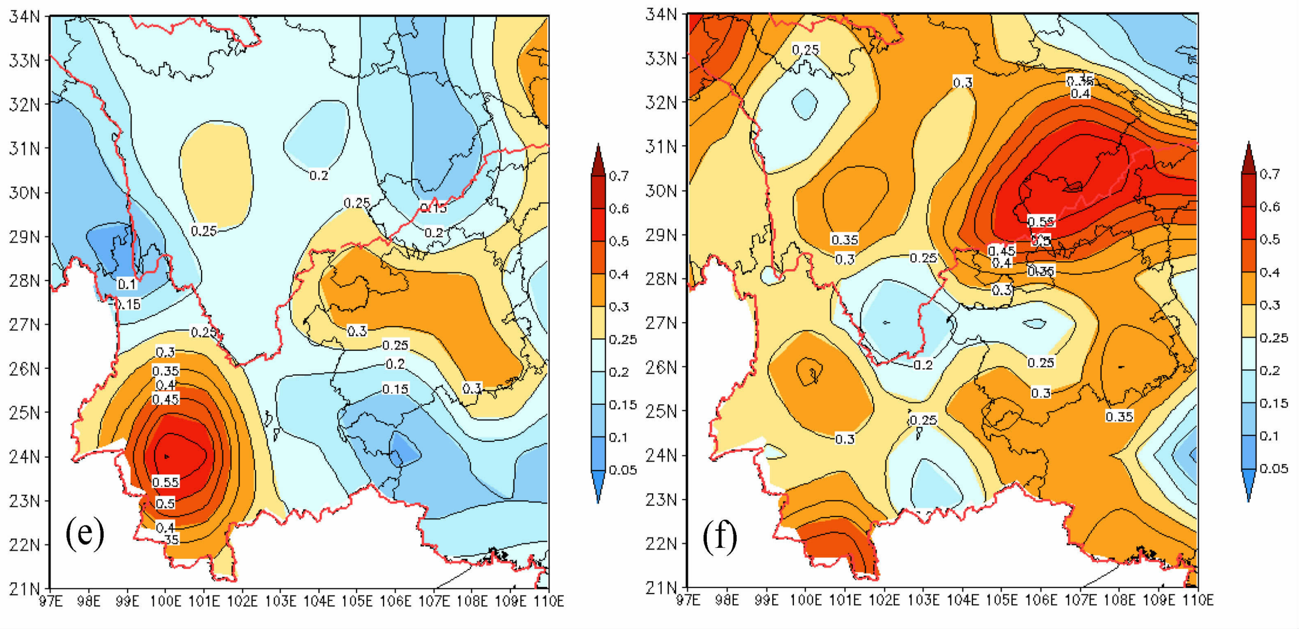


Figure 9. The spatial distribution of the assessment scores of ACC, RMSE and MAE of the PLS model and the MGF improved PSO-PLS model. (a) MAE (PLS) (b) MAE (the MGF improved PSO-PLS) (c) RMSE (PLS) (d) RMSE (the MGF improved PSO-PLS) (e) ACC (PLS) (f) ACC (the MGF improved PSO-PLS).

Table 7. Cross-verification of prediction skills for regional rainfall of flood season over southwestern China through the MGF improved PSO-PLS regression method.

Alternative factors	13	15	17
10% test samples	73.34	72.52	74.05
20% test samples	71.62	70.84	71.90
30% test samples	70.84	71.50	70.76

Table 7 indicates that PS scores of the MGF improved PSO-PLS model varies with the proportion of random sampling and the number of initial particle population. The results show that the PS score slowly rises with an increase in the number of modeling samples and a decrease in the number of the test samples in general. Its variation range is always less than 4%. Meanwhile, there is no significant effect of the initial number of alternative predictors in the PSO algorithm, which indicates that the PSO algorithm is objective and stable. The abovementioned analysis shows that the prediction performance of the model is reliable and acceptable.

7 SUMMARY

For the forecasting of PreFS, a method is proposed to automatically select the optimal factor combination for the PLS regression model, namely, the PSO-PLS regression method. Using the PSO-PLS regression method, a forecast experiment for the 36-station and whole region PreFSs is conducted. The difference between the PSO-PLS regression model and the PLS model is compared. The periodic extension sequences are constructed through MGF and the most correlative ones are considered to enrich the predictors. In this way, the problem of the low forecast accuracy of the model in anomalous precipitation year is improved. The stability of the model is also analyzed. The following conclusions are obtained:

(1) The model fitting results and forecast results of PreFS by the PSO-PLS regression method and PLS regression method are compared. It is showed that the fitting results of two methods are considerable. The forecast results showed that PSO-PLS is superior to the PLS method in terms of RMSE, MAE, ACC, and PS score. The accuracy of the model prediction results of the PLS regression method are significantly improved with the PSO algorithm.

(2) Because the number of factors showing good correlation with the PreFS is insufficient in the PSO-PLS regression model, the precipitation sequences extended by MGF are added as predictive factors to improve the model. For the abnormal precipitation year 2011, the PS score of the improved model is 75, much higher than that of the other two models. It is proved that accuracy is effectively improved by using the improved PSO-PLS model to predict the PreFS in anomalous year.

(3) The stability of the PSO-PLS model is verified by the random cross validation. It shows that with an increase in the number of modeling samples, the PS score slowly increases but the variation range of the PS score is always less than 4% . The impact on the prediction results is not significant when changing the

initial number of alternative factors in the PSO algorithm. This shows that the PSO algorithm is objective and stable. The PSO-PLS precipitation prediction model has good stability and forecast ability; it has good application perspective.

REFERENCES:

[1] HUA Wei, FAN Guang-zhou, ZHOU Ding-wen, et al. Characteristics of winter and spring vegetation variation over Tibetan plateau and its influence on summer precipitation of southwest China [J]. Sci Meteor Sinica, 2008, 28(4): 363-369 (in Chinese).

[2] SHEN Yong-ping, WANG Guo-ya. Key findings and assessment results of IPCC WGI Fifth Assessment Report [J]. J Glaciol & Geocryol, 2013, 35 (5): 1068-1076 (in Chinese).

[3] TRENBERTH K E. Uncertainty in hurricanes and global warming [J]. Science, 2005, 308(5729): 1753-1754.

[4] KHARIN V V, ZWIERS F W. Estimating extremes in transient climate change simulations [J]. J Climate, 2005, 18(8): 1156-1173.

[5] ZHONG Zhong, HU Yi-jia, MIN Jin-zhong. Sensitivity of interannual and interdecadal precipitation variability over China to spatial scale [J]. Chin J Geophys, 2007, 50(5): 1152-1159.

[6] WOLD S, ALBANO C, DUNN W J, et al. Modelling data tables by principal components and PLS: Class patterns and quantitative predictive relations [J]. Analusis, 1983, 12 (10): 477-485.

[7] WOLD S, ALBANO C, DUNN W J, et al. Pattern recognition: finding and using regularities in multivariate data food research, how to relate sets of measurements or observations to each other [C]//Food research and data analysis: proceedings from the IUFoST Symposium. Oslo, Norway, 1983.

[8] HÖSKULDSSON A. PLS regression methods [J]. J Chemometr, 1988, 2(3): 211-228.

[9] GELADI P, KOWALSKI B. Partial least squares regression: a tutorial [J]. Anal Chem Acta, 1986, 185(1): l-17.

[10] ERTAC M, FIRUZAN E, SOLUM S. Forecasting Istanbul monthly temperature by multivariate partial least square [J]. Theor Appl Climatol, 2015, 121(1-2): 253-265.

[11] KULKARNI M A, ACHARYA N, KAR S C, et al. Probabilistic prediction of Indian summer monsoon rainfall using global climate models [J]. Theor Appl Climatol, 2012, 107(3): 441-450.

[12] WANG Qi-guang, SU Hai-jin, ZHI Rong, et al. The analogy and predictability of the forecasting model error for the precipitation over the mid-lower reaches of the Yangtze River in summer [J]. Acta Phys Sinica, 2014, 63 (11): 119202-1-119202-9 (in Chinese).

[13] MASSY W F. Principal component regression in exploratory statistical research [J]. J Amer Stat Assoc, 1965, 60(309): 234-256.

[14] HOERL A E, KENNARD R W. Ridge regression: biased estimation for non-orthogonal problems [J]. Technometr, 2000, 42(1): 55-88.

[15] WEBSTER J T. Latent root regression analysis [J]. Technometr, 1974, 16(4): 513-522.

[16] FURNIVAL G M. All possible regressions with less computation [J]. Technometr, 1971, 13(2): 403-408.

[17] SEHATZOFF M, TSAO R, FIENBERG S. Efficient calculation of all possible regressions [J]. Technometr, 1968, 10(4): 769-779.

[18] WOLD S, RUHE A, WOLD H, et al. The collinearity problem in linear regression, the partial least squares (PLS) approach to generalized inverses [J]. SIAM J Sci Stat Comput, 1984, 5(3): 735-743.

[19] ABDI H, WILLIAMS L J. Partial Least Squares Methods: partial least squares correlation and partial least square regression [M]// REISFELD B, MAYENO A (eds), Methods in Molecular Biology. New York: Humana Press, 2013: 549-579.

[20] YAN Hong-ming, LI Qin-quan, SUN Chen-hu, et al. Criterion for determining the onset and end of the rainy season in Southwest China [J]. Chin J Atmos Sci, 2013, 37(5): 1111-1128 (in Chinese).

[21] JIN L, ZHU J, HUANG Y, et al. A nonlinear statistical ensemble model for short-range rainfall prediction [J]. Theor & Appl Climatol, 2015, 119(3–4): 791-807.

[22] KENNEDY J, EBERHART R C. Particle swarm optimization [C]//Proceeding of IEEE International Conference on Neural Networks. Perth, Australia, 1995: 1942-1948.

[23] EBERHART R C, KERMEDY J. A new optimizer using particle sworn theory [C]//Proceeding of the Sixth International Symposium on Micro Machine and Human Science. Nagoya, Japan, 1995: 39-43.

[24] EBERHART R C, SHI Y. Particle swarm optimization: developments, applications and resources [C] //Proceedings of the IEEE Congress on evolutionary computation. Piscataway, NJ: IEEE Service Center, 2001: 81-86.

[25] KENNEDY J, EBERHART R C. A discrete binary version of the particle swarm algorithm [C]//Proceedings of the World Multiconference on Systemics, Cybernetics and Informatics. Piscataway, NJ: IEEE Service Center, 1997: 4104-4109.

[26] LIU Jian-hua. The research of basic theory and improvement on particle swarm optimization [D]. Changsha: Central South University, 2009: 46-89 (in Chinese).

[27] KENNEDY J. The behavior of particles [C]//7th Annual Conference on evolutionary programming. San Diego, USA, 1998: 998-1105.

[28] YING L L, EL-SALEH A A, LOO J, et al. Performance investigation on binary particle swarm optimization for global optimization [C]//International Conference on Practical Applications of Agents and Multi-Agent Systems. Springer International Publishing, 2015: 142-154.

[29] LIU Z X, HUA L. Parameter setting and experimental analysis of the random number in particle swarm optimization algorithm [J]. Ctrl Theor & Appl, 2010, 27 (11): 1489-1496.

[30] WOLD S, KETTANEH-WOLD N, SKAGERBERG B. Non-linear PLS modeling [J]. Chemom Intell Lab Syst, 1989, 7(1): 53-65.

[31] WANG Hui-wen. Partial Least Squares Regression Method and Its Application [M]. Beijing: National Defense Industry Press, 1999: 46-78 (in Chinese).

[32] WEI Feng-yin, CAO Hong-xin. The new scheme of establishing long term forecast model and its application

[J]. Chin Sci Bull, 1990, 35(10): 777-780 (in Chinese).

[33] WEI Feng-yin. Modern Climate Statistics Diagnosis and Prediction Technology [M]. Beijing: Chin Meteor Press, 2007: 67-80 (in Chinese).

[34] TONG J C. Cross-Validation [M]. Springer, New York, 2013: 35-64.

[35] XIONG Qiu-fen, GU Yong-gang, WANG Li. Application of SVM method to cloud amount forecast [J]. Meteor Mon, 2007, 33(5): 20-26 (in Chinese).

Citation: WANG Zhi-yi, HU Bang-hui, YANG Xie-qun, et al. Prediction of flood season precipitation in Southwest China based on improved PSO-PLS [J]. J Trop Meteor, 2018, 24(2): 163-175.