Water Resour Manage
DOI 10.1007/s11269-016-1288-8

Monthly Rainfall Forecasting Using EEMD-SVR Based on Phase-Space Reconstruction

Qi Ouyang [1,2] · **Wenxi Lu** [1,2] · **Xin Xin** [1,2] · **Yu Zhang** [1,2] · **Weiguo Cheng** [1,2] · **Ting Yu** [1,2]

Received: 4 June 2015 / Accepted: 9 March 2016

Abstract Rainfall links atmospheric and surficial processes and is one of the most important hydrologic variables. We apply support vector regression (SVR), which has a high generalization capability, to construct a rainfall forecasting model. Before construction of the model, a self-adaptive data analysis methodology called ensemble empirical mode decomposition (EEMD) is used to preprocess a rainfall data series. In addition, the phase-space reconstruction method is implemented to design input vectors for the forecasting model. The proposed hybrid model is applied to forecast the monthly rainfall at a weather station in Changchun, China as a case study. To demonstrate the capacity of the proposed hybrid model, a typical three-layer feed-forward artificial neural network model, an auto-regressive integrated moving average model, and a support vector regression model are constructed. Predictive performance of the models is evaluated based on normalized mean squared error (NMSE), mean absolute percent error (MAPE), Nash–Sutcliffe efficiency (NSE), and the coefficient of correlation (CC). Results indicate that the proposed hybrid model has the lowest NMSE and MAPE values of 0.10 and 14.90, respectively, and the highest NSE and CC values of 0.91 and 0.83, respectively, during the validation period. We conclude that the proposed hybrid model is feasible for monthly rainfall forecast and is better than the models currently in common use.

Keywords Rainfall forecasting · Support vector regression · Ensemble empirical mode decomposition · Phase-space reconstruction

✉ Xin Xin
xxxx@jlu.edu.cn

[1] Key Laboratory of Groundwater Resources and Environment, Ministry of Education, Jilin University, Changchun 130021, China

[2] College of Environment and Resources, Jilin University, Changchun 130021, China

Published online: 17 March 2016

1 Introduction

Rainfall is one of the most important aspects of the hydrologic cycle because it links atmospheric and surficial processes. Over the past few decades, rainfall forecasting has been of great concern to researchers (Maheswaran and Khosa 2014; Wu et al. 2015; George et al. 2016). Broadly, there are two main approaches to rainfall forecasting: (1) a knowledge-driven, physically-based modeling approach; and (2) a data-driven, empirically-based or 'black-box' modeling approach (Hong and Pai 2006; Solomatine and Ostfeld 2008). The former approach is based on the physical mechanisms of hydrologic processes, and is usually based on the characteristics and scientific understanding of a specific catchment, whereas the latter approach is designed to identify relationships between input and output without considering the internal structure of physical processes (Chau et al. 2005; He et al. 2014).

A knowledge-driven modeling approach can involve detailed description of the mechanisms of hydrological processes. However, the data required (e.g., temperature, pressure, humidity) are extensive and sometimes unavailable (Hong 2008). Furthermore, it is challenging to extend a particular knowledge-driven model to even a slightly different research area (Sivakumar et al. 2002). The data-driven modeling approach, which uses time series data, is relatively simple and compatible across regions, and has been widely employed in the forecast of hydrologic variables (Kaur and Jothiprakash 2013). The artificial neural network (ANN) and Box–Jenkins methods (Box and Jenkins 1970), which include an auto-regressive moving average (ARMA) model, an auto-regressive integrated moving average (ARIMA) model, an auto-regressive (AR) model, and a moving average (MA) model, have been the most widely used data-driven models for the last few decades (Paolo et al. 1993; Valverde Ramírez et al. 2005; Chua and Wong 2011; Farajzadeh et al. 2014). The results of the previous work suggest that the ANN models perform better when variables are nonlinear, while the Box–Jenkins methods are more successful with linear variables.

Support vector machines (SVM), developed by Vapnik (1995), are learning machines based on statistical learning theory that adopt the structural risk minimization principle rather than the empirical risk minimization principle (the principle followed by ANN). According to previous studies (e.g. Sivapragasam et al. 2001; Liong and Sivapragasam 2002), SVM can better solve problems of small sample size, overlearning, nonlinearity, high dimensionality, and local minima than ANN can, and has high generalization capability (Wang et al. 2013). The regression model of SVM, called support vector regression (SVR), has been successfully employed to solve forecasting problems with hydrologic variables, such as rainfall in general (Hong and Pai 2006; Feng et al. 2015), typhoon rainfall (Lin et al. 2009), groundwater level (Suryanarayana et al. 2014), rainfall runoff (Wang et al. 2013), real-time daily flow (Maheswaran and Khosa 2013), lake water level (Khan and Coulibaly 2006), and riverine suspended sediment load (Nourani et al. 2016). However, there are problems with the employment of SVR, primarily in two aspects. Firstly, an assumption of this method is the stationarity of the original data. Unfortunately, in reality, time series data rarely adhere to this assumption due to fluctuation and intrinsic complexity (Hu et al. 2013a). Therefore, for more accurate forecasting results, it is crucial to apply suitable data preprocessing before prediction. Secondly, the design of the input vector is important for the time series prediction engine, but it is usually defined in an arbitrary way based mainly on experience.

For data preprocessing, wavelet transform (WT) has been found to be a good choice in recent years (He et al. 2014; Suryanarayana et al. 2014; Feng et al. 2015). However, recent studies suggest that WT suffers from certain drawbacks (Zhang and Zhou 2013). Empirical mode decomposition (EMD), first introduced by Huang et al. (1998), is a technique that offers a different method for data preprocessing. Based on local characteristic time scales of a signal, EMD can self-adaptively decompose a complex signal into a series of intrinsic mode functions (IMFs) and one residue and remove the noise. IMFs represent the natural oscillatory mode embedded in the signal; each is simple and has its corresponding physical meaning and frequency, which allow for better understanding of the mechanics behind the signal. However, the frequent occurrence of mode mixing hampers its application. Wu and Huang (2009) added finite white noise to signals to overcome this drawback of EMD, and called this method ensemble empirical mode decomposition (EEMD). EEMD has captured scholars' attention globally in a wide range of fields; e.g., Bao et al. (2012) combined EEMD and SVM to forecast air passenger traffic; Wang et al. (2012) compared application of EMD and EEMD on time–frequency analysis of seismic signals, and demonstrated that the time–frequency spectrum obtained by EEMD more realistically reflects real geology than that obtained by EMD; Hu et al. (2013a) applied a hybrid EEDM and SVM approach to forecast wind speed time series data, and the results indicated an observable improvement to forecasting validity; Wang et al. (2013) implemented a PSO-SVM-EEMD model to forecast annual rainfall runoff, and found that this methodology can significantly improve rainfall-runoff forecasting at the studied station.

Numerous studies in recent years have confirmed the existence of chaotic behavior in hydrologic processes, including runoff, rainfall, floods, lake level, and evaporation (e.g., Damle and Yalcin 2007; Dhanya and Nagesh Kumar 2011a, b; Hu et al. 2013b; Khatibi et al. 2014). Based on chaos theory, a random-seeming series of chaotic behaviors can be attributed to deterministic rules (Ng et al. 2007). Under Takens' embedding theorem (Takens 1981), phase-space reconstruction can provide a favorable solution to fully uncover the underlying dynamics of a deterministic chaotic system by building an m-dimensional space. With this method, the design of the input vector can be solved. Kouhi et al. (2014) and Baydaroğlu and Koçak (2014) applied this approach to prepare input data for time series prediction and were successful.

The present study seeks to address both problems in rainfall time series prediction, i.e., the non-stationarity of rainfall time series data, and the design of input vectors. Firstly, instead of WT, a relatively new data preprocessing method, EEMD, is employed to decompose the original rainfall data into a series of IMFs and one residue, thereby transforming non-stationarity into stationarity. Subsequently, the phase-space reconstruction method is utilized to build an m-dimensional space to recover the dynamics of each IMF to prepare the input data for prediction. Based on the above techniques, SVR, a promising nonlinear regression learning machine, is employed to forecast future values of each IMF and residue. These results can be assembled to determine the final forecast results of rainfall.

The present study is the first to combine SVR, EEMD, and the phase-space reconstruction method.

2 Study Area and Data

Changchun, the capital city of Jilin Province in northeastern China, lies between the latitudes of 43°05' N and 45°15' N, and between longitudes of 124°18' E and 127°05' E, at an elevation between 250 m and 350 m. Located on the Songliao Plain, the hinterland of the Northeast China Plain on the east coast of Eurasia, Changchun is the natural geographical center of northeastern China, which has a total area of 20,604 km^2 and a total population of 790 million. The Yitong River runs through the city from south to north. A map of Jilin Province is shown in Fig. 1. This region lies in the north temperate continental monsoon climate zone, and rainfall is distributed unevenly throughout the year. Average annual rainfall ranges from 522 to 615 mm, and summer precipitation accounts for more than 60 % of that amount. Observed monthly rainfall from the Changchun weather station from 1951 to 2013 was selected as the dataset for this study.

3 Methodology

3.1 SVR

SVM is a novel learning machine with a high generalization capability. SVR, a subcategory of SVM, is proposed to solve the regression problem present in SVR. The concept of SVR is to

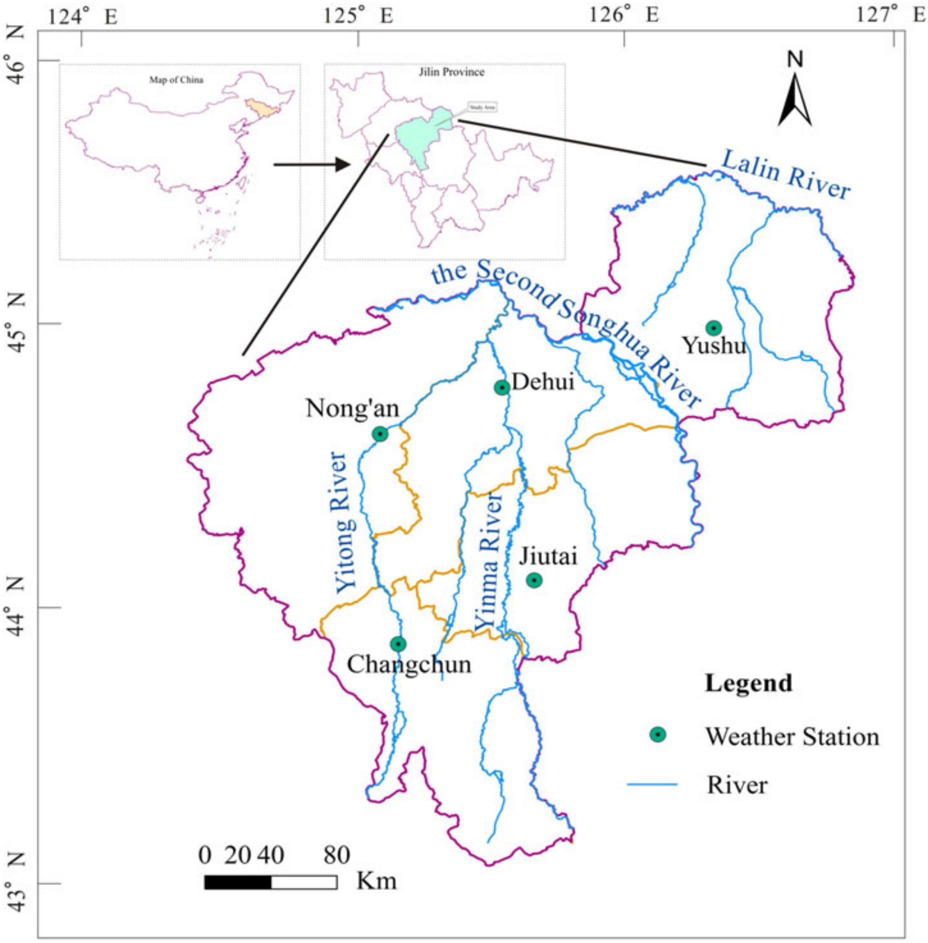


Fig. 1 Location map of Changchun and the surrounding area

map nonlinearly the original data x into a high-dimensional feature space (or even an infinite dimensional space), and then to perform a linear regression in the feature space (Wang et al. 2013). Given a set of training data $(x_1, y_1),\ldots, (x_l, y_l)$ where $x_i \in R^n$ is the input vector and $y_i \in R$ for $i=1, \ldots, l$, represents the respective target value (or output value), and l denotes the number of elements in the training data set. SVR estimates the output at a prediction point x_p as

$$\begin{aligned}\hat{y}(x_p) &= \sum_{i=1}^{l} \left(\alpha_i - \alpha_i^{*}\right) K\left(x_p \cdot x_i\right) + b \\ K\left(x_p, x_i\right) &= \left[\varphi\left(x_p\right) \ , \varphi(x_i)\right]\end{aligned} \tag{1}$$

where α and α^* are the dual Lagrange multipliers, b is the bias term, and $K(x_p, x_i)$ is the kernel function. In general, there are several types of kernel function, namely linear, polynomial and radial basis function (RBF). RBF has become the most popular of these and is adopted in the present study as follows:

$$K(x, x_i) = \exp\left(-\gamma \|x - x_i\|^2\right) \tag{2}$$

where γ is an unknown parameter.

3.2 EEMD

The predecessor of EEMD is EMD, which was first introduced by Huang et al. (1998). The essence of this method is to process data with axial symmetry and decompose it into a series of IMFs in descending order by signal frequency. An IMF must satisfy the following conditions:

a) In the whole data set, the number of extrema and the number of zero-crossings must either equal or differ by at most one.
b) At any point, the mean value of the envelope defined by local maxima and the envelope defined by the local minima is zero.

IMFs can be extracted from the data series $X(t)$ according to a so-called sifting process (Huang *et al.* 1998). After the process of EMD, the original time series data $X(t)$ can be expressed as a sum of IMFs and one residue:

$$X(t) = \sum_{i=1}^{n} C_i + r_n \tag{3}$$

where n is the number of IMFs, C_i represents the ith generated IMF, and r_n denotes the final residue that represents the overall trend of the data series $X(t)$.

Despite the wide acceptance of EMD, the frequent occurrence of mode mixing seems to be a significant drawback. This problem implies that either a single IMF consisting of signals of widely disparate scales, or a signal of the a specific scale that resides in different IMF components, and intermittency of the analyzed signal is often the result (Wang et al. 2012). To overcome this problem, Wu and Huang (2009) developed EEMD, which adds white noise into the original data, and thereby signals of different scales can be automatically assigned to proper scales of reference established by the white noise (Wang et al. 2012), which reduces the occurrence of mode mixing.

3.3 Phase-Space Reconstruction

The phase-space reconstruction method can fully uncover the underlying dynamics of a deterministic chaotic system by reconstructing the phase space, which provides a simplified, multi-dimensional representation of a single-dimensional nonlinear time series. By this method, for a scalar time series X_i where i =1, 2, …, N, the dynamics can be fully embedded in m-dimensional phase space wherein the components of each state vector Y_j are defined through the delay coordinates:

$$Y_j = \left(X_j,\ X_{j+\tau},\ X_{j+2\tau},\ \dots,\ X_{j+(m-1)\tau}\right) \tag{4}$$

where $j = 1, 2, \dots, N-(m-1)\tau/\Delta t$; τ is the delay time, which is the average length of memory of the chaotic system; m is the embedding dimension, which can be considered the minimum number of state variables required to describe the system, and Δt represents sampling time. Phase-space reconstruction in a dimension m allows one to interpret the underlying dynamics in the form of an m-dimensional map F_T, that is,

$$Y_{j+T} = F_T\left(Y_j\right) \tag{5}$$

where Y_{j+T} is the vector describing the state of the system at time $j+T$ (the future state), and T refers to the lead time. Thus, we can design the input data for regression modeling. Here, the m-dimensional map F_T is constructed through SVR.

The determination of the two parameters, i.e., τ and m, is crucial for the correct identification of hidden dynamics, and there are a variety of ways to estimate them. The autocorrelation method and mutual information method (Frazer and Swinney 1986) are frequently-used approaches to identify τ; the correlation dimension method (Grassberger and Procaccia 1983), the false nearest-neighbor algorithm (FNN) (Kennel et al. 1992), and the Cao method (Cao 1997) are popular approaches to identify m. According to previous studies (e.g., Hu et al. 2013b; Guo et al. 2014), the mutual information method and the Cao method are the most favorable of these techniques and are adopted for this study.

After estimation of m and τ, the following step is to identify the presence of chaotic behavior. The Lyapunov exponent λ is the most commonly used indicator of chaotic behavior. To be determined chaotic, the largest Lyapunov exponent, $\lambda_{\max}$, must be positive. There are a multitude of approaches to calculate λ, such as the p-norm method, the Wolf method, the Jacobian method, and the small data sets method (Rosenstein et al. 1993). Because the small data sets method has relatively high computational efficiency and accuracy (Rosenstein et al. 1993, Hu et al. 2013b), it is employed here to compute $\lambda_{\max}$.

4 Model Constructions

For this study, first, EEMD is employed as a data preprocessing method for monthly rainfall data taken at the Changchun weather station from 1951 to 2013. Secondly, for the design of input data, phase-space reconstruction is used, and the delay time τ and embedding dimension m of each IMF (and the residue) are determined using the mutual information method and the Cao method respectively. Subsequently, we construct the nonlinear regression model of each IMF (and the residue) independently by SVR to forecast future values. Finally, by assembling the forecasted values of each IMF (and the residue) into an ensemble result, we attain the predictions of monthly rainfall values. Figure 2 is a flow chart that illustrates the model construction procedure.

4.1 Data Preprocessing With EEMD

Before using EEMD there are two important parameters to set, i.e., the number of the ensemble and the amplitude of the added white noise. Wu and Huang (2009) established a statistical rule to control the effect of noise:

$$e_n = \frac{\varepsilon}{\sqrt{N}} \tag{6}$$

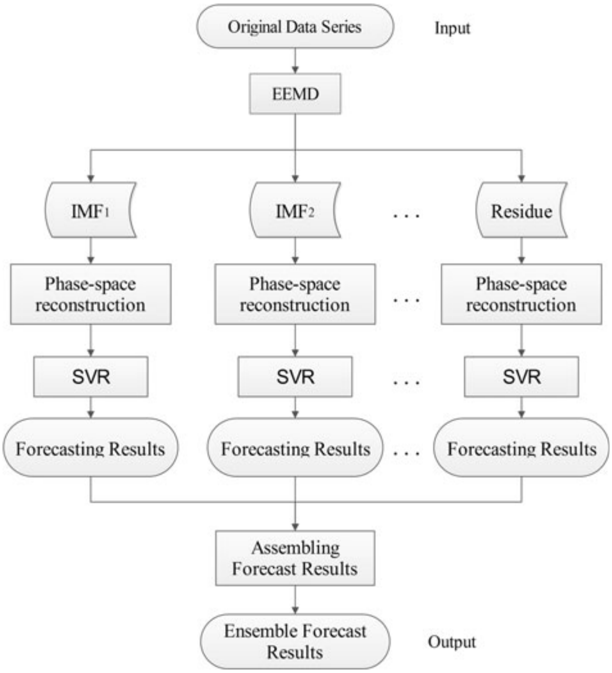
where N is the number of ensemble members, ε represents the amplitude of the added noise, and e_n is the final standard deviation, which is defined as the difference between the input signal and the corresponding IMFs (Wu and Huang 2009). Here, as in previous studies (e.g. Wu and Huang 2009; Wang *et al.* 2013), N and e_n are set as 100 and 0.2, respectively.

Decomposition results are shown in Fig. 3. There are eight independent IMF compositions and one residue.

4.2 Phase-Space Reconstruction

For this study, the mutual information method and the Cao method are used to determine the parameters τ and m respectively. The determination results for IMF_1 and IMF_2 are shown in Figs. 4 and 5 as examples. Table 1 shows the determination results for IMF_1, IMF_2,…, IMF_8, residue, and the original data.

Fig. 2 Flow chart of the model construction procedure


Original Data Series
Input
EEMD
IMF1
IMF2
. . .
Residue
Phase-space reconstruction
Phase-space reconstruction
Phase-space reconstruction
SVR
SVR
SVR
Forecasting Results
Forecasting Results
Forecasting Results
Assembling Forecast Results
Ensemble Forecast Results
Output

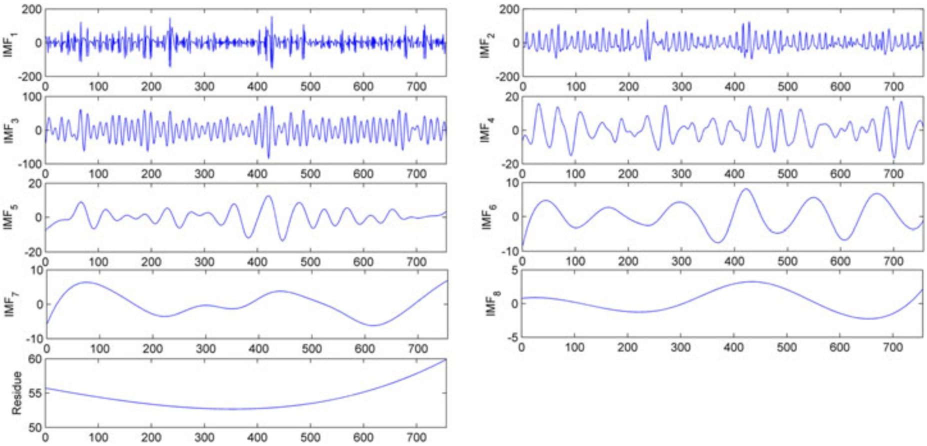


Fig. 3 Decomposition by EEMD of monthly rainfall from January 1951 to December 2013 at the Changchun weather station

After determination of the above two parameters, the small data sets method is applied to calculate $\lambda_{\max}$ of each IMF, residue, and the original data. The computed $\lambda_{\max}$ for IMF$_1$, IMF$_2$,…, IMF$_8$, residue and the original data are also shown in Table 1. All of the largest Lyapunov exponents are positive, which indicates chaos. The one-dimensional data series can be assembled into the following m-dimension matrix:

$$X = \begin{bmatrix} x_1 & x_{1+\tau} & \dots & x_{1+(m-1)\tau} \\ x_2 & x_{2+\tau} & \dots & x_{2+(m-1)\tau} \\ \dots & \dots & \dots & \dots \\ x_{n-(m-1)\tau} & x_{n-(m-1)\tau+1} & \dots & x_{n-1} \end{bmatrix}, \quad Y = \begin{bmatrix} x_{2+(m-1)\tau} \\ x_{3+(m-1)\tau} \\ \dots \\ x_n \end{bmatrix} \tag{7}$$

where X is the input vector and Y is the output vector. The lead time is set to be 1 month.

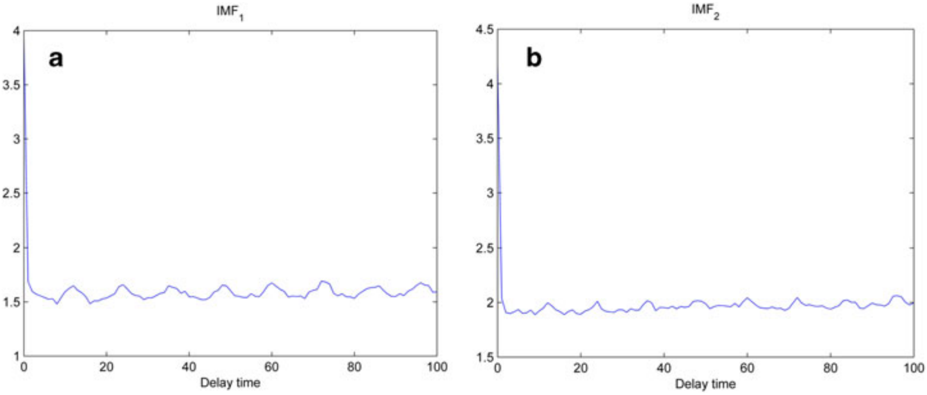


Fig. 4 Mutual information results plot of IMF$_1$ **a** and IMF$_2$ **b** for the determination of time delay

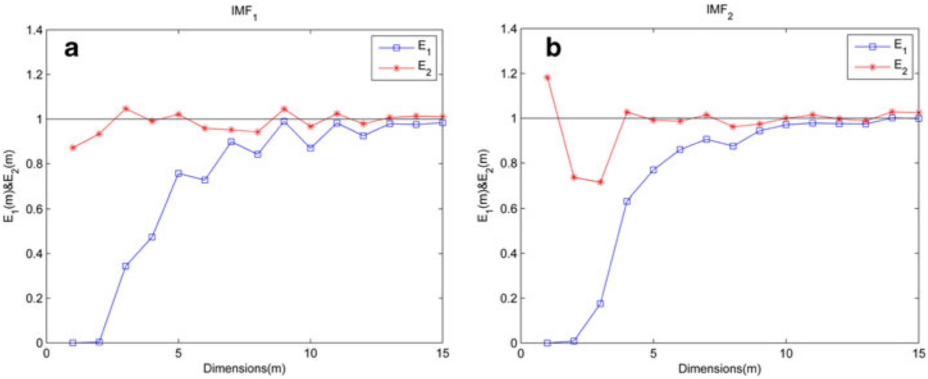


Fig. 5 Cao method results plot of IMF$_1$ **a** and IMF$_2$ **b** for the determination of embedding dimension

4.3 SVR Model

Here, the SVR model is used independently to forecast the future values of each IMF (and the residue). Data from the first 50 years (1951–2000, 600 samples in total) are used as the training dataset, while data from the latter 13 years (2001–2013, 156 samples in total) are used as a validating dataset. With the training dataset, the grid search method is employed to find optimal parameters of. C and γ for SVR, and the insensitive loss function ε is set to "0.01" based on prior experience. Then, by assembling the outcomes of each SVR model, we can obtain the ensemble forecast rainfall value. The optimization results of the SVR parameters are shown in Table 1.

5 Performance Evaluation

In the present study, four main criteria are used to measure the models' forecasting performance. The normalized mean squared error (NMSE) and the mean absolute percent error (MAPE), given by Eq. (8) and Eq. (9) respectively, are used to measure the accuracy of forecasting. The smaller the value of NMSE and MAPE, the more accurate the model

Table 1 Parameter results for phase-space reconstruction and SVR model

Data	τ	m	$\lambda_{\max}$	γ	C
IMF$_1$	6	8	0.100	0.8123	4.2871
IMF$_2$	3	11	0.104	0.4353	2.0000
IMF$_3$	3	6	0.036	1.4142	4.0000
IMF$_4$	5	6	0.035	0.5359	9.8492
IMF$_5$	6	10	0.013	0.5000	27.8576
IMF$_6$	13	6	0.020	0.2679	51.9842
IMF$_7$	9	5	0.024	3.0314	294.0668
IMF$_8$	4	5	0.017	5.6569	415.8732
Residue	3	3	0.014	0.5359	724.0773
Original Data	5	3	0.0003	2.1435	0.2500

determined to be. In addition, the efficiency of the model is measured in terms of the Nash–Sutcliffe efficiency coefficient (NSE) and the coefficient of correlation (CC), given by Eq. (10) and Eq. (11) respectively. NSE is a frequently used index for evaluating the predictive ability of hydrological models. The higher value of NSE (maximum value is 1), the higher the model's forecast power is. Similarly, a model with a higher value of CC (to a maximum value of 1) can better capture the average change tendency of the cumulative data series (Hong 2008), which means a high degree of collinearity. The questions for these evaluation methods are as follows:

$$\mathrm{NMSE} = \frac{1}{n\delta^2}\sum_{i=1}^{n}(a_i - f_i)^2 \tag{8}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{a_i - f_i}{a_i}\right| \tag{9}$$

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}(a_i - f_i)^2}{\sum_{i=1}^{n}\left(a_i - \overline{a}\right)^2} \tag{10}$$

$$\mathrm{CC} = \frac{\sum_{i=1}^{n}\left(a_i - \overline{a}\right)\left(f_i - \overline{f}\right)}{\sqrt{\sum_{i=1}^{n}\left(a_i - \overline{a}\right)^2 * \sum_{i=1}^{n}\left(f_i - \overline{f}\right)^2}} \tag{11}$$

where $\delta^2 = \frac{1}{n-1}\sum_{i=1}^{n}(a_i - \overline{a})^2$, and n is the number of forecasting periods; a_i and f_i denote the actual and forecast rainfall values respectively; and $\bar{a}$ and $\overline{f}$ represent the actual and forecast mean rainfall values respectively.

6 Results and Discussions

In order to evaluate the advantage of the proposed hybrid model (I), a typical three-layer feed-forward ANN model (II) and an ARIMA model (III) are constructed as benchmark models. The two benchmark models also use EEMD to pre-process the original monthly rainfall data and adopt the phase-space reconstruction method to design input vectors, and the training and validating data sets are identical across all models. Because the only difference between models I, II and III is which forecasting technique is used, SVR, ANN, or ARIMA, by comparing their results, we can ascertain which is the most accurate. Additionally, a SVR model is constructed to use the original monthly rainfall series (IV), with phase-space reconstruction applied but EEMD omitted, in order to evaluate the effect of EEMD on forecast accuracy by applying only phase-space reconstruction and comparing its results with model I.

Table 2 illustrates the evaluation results for the four different models in terms of the four indices mentioned above. These results indicate the following: 1) the proposed hybrid model (I) has the lowest NMSE and MAPE value and the highest NSE and CC value, at both the training and validation stage, which demonstrates that the proposed model outperforms the other three models in forecasting monthly rainfall. 2) By comparing the results for models I and II, we observe that relative to model II, for the proposed hybrid model (I), NMSE and MAPE decreased by 87.04 % and 45.76 % respectively in the training stage, and by 89.89 % and 42.40 % respectively in the validation stage. CC and NSE were improved by 9.47 % and 100%respectively in the training stage, and by 37.88 % and 84 % respectively in the validation stage. Similarly, by comparing model I and III, we observe that the relative to model III, the proposed hybrid model (I) presents decreases of 75.76 % and 69.05 % in NMSE and MAPE respectively in the training stage, and 73.68 % and 68.81 in the validation stage; and was improved by 6.74 % and 19.48 % in CC and NSE respectively in the training stage, and by 10.98 % and 33.87 % in the validation stage. These results show that the SVR model outperforms the commonly used nonlinear regression models ANN and ARIMA. 3) By comparing models I and IV, we observe that the use of EEMD was responsible for a 68.18 % and 26.11 % reduction in NMSE and MAPE respectively in the training stage and a 67.74 % and 35.27 % reduction in the same in the validation stage. Improvements in the forecast results represented by CC and NSE values were approximately 5.55 % and 10.84 % respectively at the training stage, and 8.79 % and 20.29 % respectively in the validation stage, respectively. These results indicate that the pre-processing method of EEMD improved the forecasting ability of the SVR model. Figure 6 shows a comparison of the four models during the validation period.

7 Conclusions

Our findings support the employment of a hybrid model, EEMD-SVR based on phase-space reconstruction, to overcome two key problems in rainfall forecasting and improve predictive accuracy. To reasonably evaluate the proposed model's performance, a typical three-layer

Table 2 Performance indices of models for rainfall forecasting in training and validating datasets

Model	Training data				Validating data			
	NMSE	MAPE (%)	CC	NSE	NMSE	MAPE (%)	CC	NSE
I	0.07	10.50	0.95	0.92	0.10	14.90	0.91	0.83
II	0.54	19.36	0.86	0.46	0.99	25.87	0.66	−0.01
III	0.29	33.93	0.89	0.77	0.38	47.78	0.82	0.62
IV	0.22	14.21	0.90	0.83	0.31	23.02	0.83	0.69

I- EEMD-SVR based on phase-space reconstruction

II- EEMD-ANN based on phase-space reconstruction

III- EEMD-ARIMA based on phase-space reconstruction

IV- SVR based on phase-space reconstruction with the original monthly rainfall data

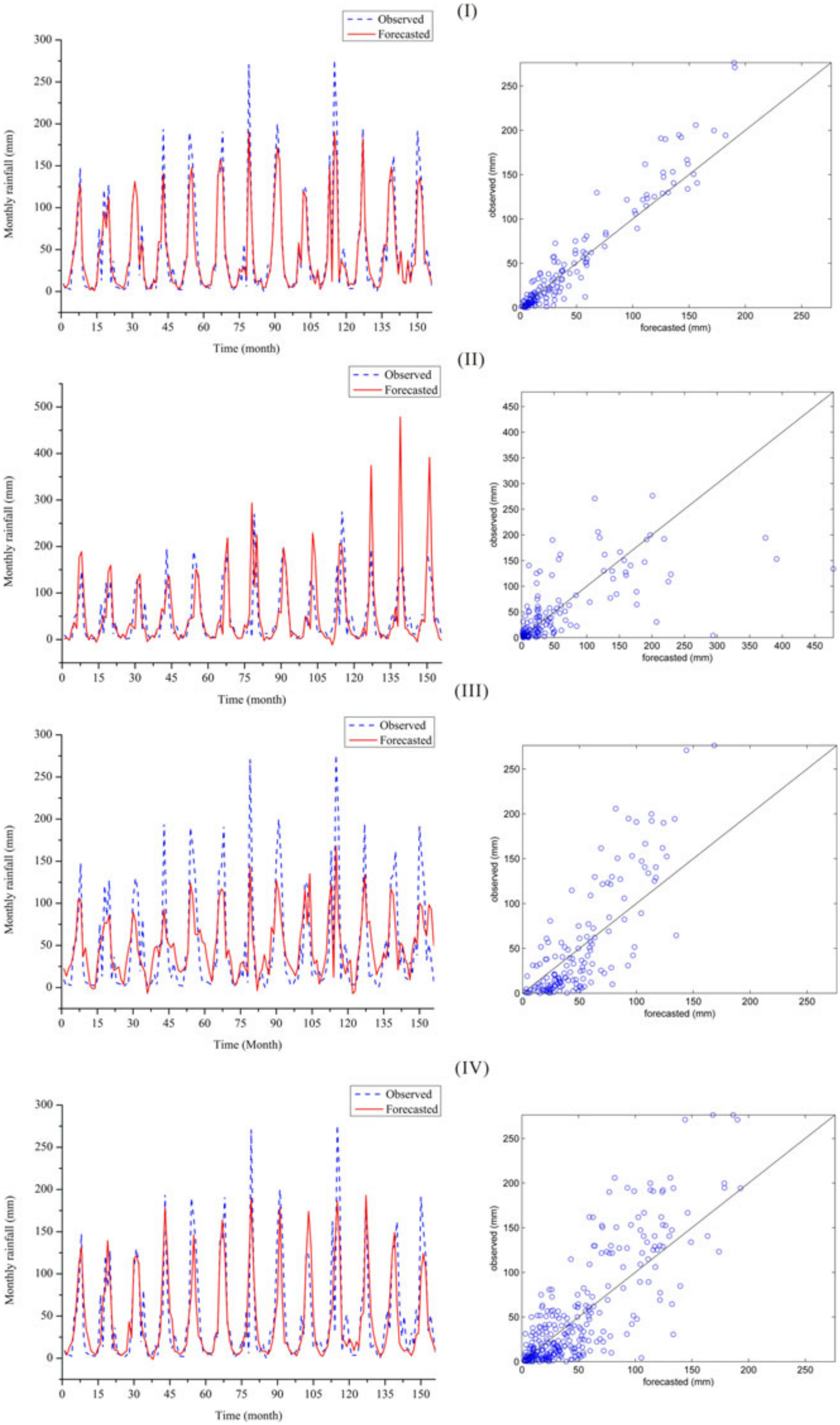


Fig. 6 Observed and forecast monthly rainfall during the validation period for four models

feed-forward ANN model, an ARIMA model, and a SVR model were constructed as benchmark models; the first two models are based on data decomposed by EEMD, the latter is based on the original data, and all the three models use phase-space reconstruction. The models we constructed are based on real monthly rainfall data from a weather station in Changchun, China. The following conclusions were reached:

1. Based on the structural risk minimization principle, SVR has better forecast results than the other assessed nonlinear regression models, which give it the advantage for rainfall forecasting.
2. By decomposing the data series into a series of independent IMFs and one residue, the data preprocessing method EEMD improved the forecast capacity of the SVR model, which indicates that EEMD is suitable for nonlinear and non-stationary hydrologic data analysis.
3. By applying phase-space reconstruction, the input vector can be designed in a certain way as an alternative to the arbitrary method.
4. The proposed hybrid model is feasible for monthly rainfall forecasting at the Changchun weather station.

Future study should focus on comparison of EEMD with other data preprocessing methods. Additionally, values of the parameters for phase-space reconstruction (i.e., delay time and embedding dimension) should be carefully determined using a variety of methods instead of one alone as is commonly done.

Acknowledgments The author would like to acknowledge the support provided by the National Natural Science Foundation of China (NO. 41372237), the Science and Technology Department of Jilin Province (NO. 20130206011SF) and the Shenyang Geological Survey Center of the China Geological Survey (NO. 1212011140027 and 12120114027401).

References

Bao Y, Xiong T, Hu Z (2012) Forecasting Air passenger traffic by support vector machines with ensemble empirical mode decomposition and slope-based method. Discret Dyn Nat Soc 2012:1–12

Baydaroğlu Ö, Koçak K (2014) SVR-based prediction of evaporation combined with chaotic approach. J Hydrol 508:356–363

Box G, Jenkins G (1970) Time series analysis forecasting and control. Holden-Day, San Francisco

Cao L (1997) Practical method for determining the minimum embedding dimension of a scalar time series. Physica D 110(1–2):43–50

Chau KW, Wu CL, Li YS (2005) Comparison of several flood forecasting models in Yangtze river. J Hydrol Eng 10(6):485–491

Chua LHC, Wong TSW (2011) Runoff forecasting for an asphalt plane by artificial neural networks and comparisons with kinematic wave and autoregressive moving average models. J Hydrol 397(3–4):191–201

Damle C, Yalcin A (2007) Flood prediction using time series data mining. J Hydrol 333(2–4):305–316

Dhanya CT, Nagesh Kumar D (2011a) Predictive uncertainty of chaotic daily streamflow using ensemble wavelet networks approach. Water Resour Res 47:28. doi:10.1029/2010wr010173

Dhanya CT, Nagesh Kumar D (2011b) Multivariate nonlinear ensemble prediction of daily chaotic rainfall with climate inputs. J Hydrol 403(3–4):292–306

Farajzadeh J, Fakheri Fard A, Lotfi S (2014) Modeling of monthly rainfall and runoff of urmia lake basin using “feed-forward neural network” and “time series analysis” model. Water Resour Indust 7–8:38–48

Feng Q, Wen X, Li J (2015) Wavelet analysis-support vector machine coupled models for monthly rainfall forecasting in arid regions. Water Resour Manag 29:1049–1065. doi:10.1007/s11269-014-0860-3

Frazer AM, Swinney HL (1986) Independent coordinates for strange attractors from mutual information. Phys Rev A 33(2):1134–1140

George J, Janaki L, Parameswaran Gomathy J (2016) Statistical downscaling using local polynomial regression for rainfall predictions – a case study. Water Resour Manag 30:183–193. doi:10.1007/s11269-015-1154-0

Grassberger P, Procaccia I (1983) Measuring the strangeness of strange attractors. Physica D 9(1–2):189–208

Guo Z, Chi D, Wu J, Zhang W (2014) A new wind speed forecasting strategy based on the chaotic time series modelling technique and the apriori algorithm. Energy Convers Manag 84:140–151

He X, Guan H, Zhang X, Simmons CT (2014) A wavelet-based multiple linear regression model for forecasting monthly rainfall. Int J Climatol 34(6):1898–1912

Hong W-C (2008) Rainfall forecasting by technological machine learning models. Appl Math Comput 200(1):41–57

Hong W-C, Pai P-F (2006) Potential assessment of the support vector regression technique in rainfall forecasting. Water Resour Manag 21(2):495–513

Hu J, Wang J, Zeng G (2013a) A hybrid forecasting approach applied to wind speed time series. Renew Energ 60:185–194

Hu Z, Zhang C, Luo G, Teng Z, Jia C (2013b) Characterizing cross-scale chaotic behaviors of the runoff time series in an inland river of central Asia. Quat Int 311:132–139

Huang NE et al (1998) The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc R Soc London, Ser A 454:903–955

Kaur H, Jothiprakash V (2013) Daily precipitation mapping and forecasting using data driven techniques. Int J Hydrol Sci Technol 3(4):364–377

Kennel MB, Brown R, Abarbanel HDI (1992) Determining embedding dimension for phase-space reconstruction using a geometirc method. Phys Rev A 45:3403–3411

Khan MS, Coulibaly P (2006) Application of support vector machine in lake water level prediction. J Hydrol Eng 11:199–205

Khatibi R et al (2014) Inter-comparison of time series models of lake levels predicted by several modeling strategies. J Hydrol 511:530–545

Kouhi S, Keynia F, Najafi Ravadanegh S (2014) A new short-term load forecast method based on neuro-evolutionary algorithm and chaotic feature selection. Int J Electr Power Energy Syst 62:862–867

Lin G-F, Chen G-R, Wu M-C, Chou Y-C (2009) Effective forecasting of hourly typhoon rainfall using support vector machines. Water Resour Res 45:8. doi:10.1029/2009WR007911

Liong SY, Sivapragasam C (2002) Flood stage forecasting with support vector machines. J Am Water Res Assoc 38(1):173–186

Maheswaran R, Khosa R (2013) Wavelets-based nonlinear model for real-time daily flow forecasting in Krishna river. J Hydroinf 15(3):1022–1041

Maheswaran R, Khosa R (2014) A wavelet-based second order nonlinear model for forecasting monthly rainfall. Water Resour Manag 28:5411–5431. doi:10.1007/s11269-014-0809-6

Ng WW, Panu US, Lennox WC (2007) Chaos based analytical techniques for daily extreme hydrological observations. J Hydrol 342(1–2):17–41

Nourani V, Alizadeh F, Roushangar K (2016) Evaluation of a Two-stage SVM and spatial statistics methods for modeling monthly river suspended sediment load. Water Resour Manag 30:393–407. doi:10.1007/s11269-015-1168-7

Paolo B, Renzo R, Luis GC, Jose DS (1993) Forecasting of short-term rainfall using ARMA models. J Hydrol 144:193–211

Rosenstein MT, Collins JJ, De Luca CJ (1993) A practical method for calculating largest Lyapunov exponents from small data sets. Physica D 65:117–134

Sivakumar B, Jayawardena AW, Fernando TMKG (2002) River flow forecasting: use of phase-space reconstruction and artificial neural networks approaches. J Hydrol 265(1–4):225–245

Sivapragasam C, Liong S-Y, Pasha MFK (2001) Rainfall and runoff forecasting with SSA-SVM approach. J Hydroinf 03(3):141–152

Solomatine DP, Ostfeld A (2008) Data-driven modelling: some past experiences and new approaches. J Hydroinf 10(1):3–22

Suryanarayana C, Sudheer C, Mahammood V, Panigrahi BK (2014) An integrated wavelet-support vector machine for groundwater level prediction in Visakhapatnam, India. Neurocomputing 14:324–335

Takens F (1981) Detecting strange attractors in turbulence, lectures notes in mathematics. Springer, New York

Valverde Ramírez MC, de Campos Velho HF, Ferreira NJ (2005) Artificial neural network technique for rainfall forecasting applied to the São Paulo region. J Hydrol 301(1–4):146–162

Vapnik V (1995) The nature of statistical learning theory. Springer, New York

Wang T, Zhang M, Yu Q, Zhang H (2012) Comparing the applications of EMD and EEMD on time–frequency analysis of seismic signal. J Appl Geophys 83:29–34

Wang W-c, Xu D-m, K-w C, Chen S (2013) Improved annual rainfall-runoff forecasting using PSO - SVM model based on EEMD. J Hydroinf 15(4):1377–1390

Wu Z, Huang NE (2009) Ensemble empirical mode decomposition: a noise-assisted data analysis method. Adv Adapt Data Analy 1:1–41

Wu J, Long J, Liu M (2015) Evolving RBF neural networks for rainfall prediction using hybrid particle swarm optimization and genetic algorithm. Neurocomputing 148:136–142

Zhang X, Zhou J (2013) Multi-fault diagnosis for rolling element bearings based on ensemble empirical mode decomposition and optimized support vector machines. Mech Syst Signal Process 41(1–2):127–140