A Hybrid LSSVM Model with Empirical Mode Decomposition and Differential Evolution for Forecasting Monthly Precipitation

LIZHI TAO, XINGUANG HE, AND RUI WANG

College of Resources and Environmental Science, Hunan Normal University, Changsha, China

(Manuscript received 10 May 2016, in final form 13 October 2016)

ABSTRACT

In this study, a hybrid least squares support vector machine (HLSSVM) model is presented for effectively forecasting monthly precipitation. The hybrid method is designed by incorporating the empirical mode decomposition (EMD) for data preprocessing, partial information (PI) algorithm for input identification, and differential evolution (DE) for model parameter optimization into least squares support vector machine (LSSVM). The HLSSVM model is examined by forecasting monthly precipitation at 138 rain gauge stations in the Yangtze River basin and compared with the LSSVM and LSSVM–DE. The LSSVM–DE is built by combining the LSSVM and DE. Two statistical measures, Nash–Sutcliffe efficiency (NSE) and relative absolute error (RAE), are employed to evaluate the performance of the models. The comparison of results shows that the LSSVM–DE gets a superior performance to LSSVM, and the HLSSVM provides the best performance among the three models for monthly precipitation forecasts. Meanwhile, it is also observed that all the models exhibit significant spatial variability in forecast performance. The prediction is most skillful in the western and northwestern regions of the basin. In contrast, the prediction skill in the eastern and southeastern regions is generally low, which shows a strong relationship with the randomness of precipitation. Compared to LSSVM and LSSVM–DE, the proposed HLSSVM model gives a more significant improvement for most of the stations in the eastern and southeastern regions with higher randomness.

1. Introduction

Reliable precipitation predictions on monthly and seasonal time scales are of great significance in hydrological research, since they can provide useful information for agricultural planning, water resource management, and associated crop insurance application (Everingham et al. 2008; Wu et al. 2010; Garbrecht et al. 2010). However, precipitation is a complex atmospheric process and is influenced by complex interactions of ocean, atmosphere, and land surface processes (Peng et al. 2014). This makes monthly and seasonal precipitation predictions particularly difficult. Consequently, it has been challenging to obtain accurate and reliable precipitation predictions (Wang et al. 2014).

In the past decades, various methods have been developed for forecasting precipitation. These approaches roughly fall into two categories: empirical and dynamical (He et al. 2015). The dynamical models based on the laws of physics such as general circulation models (GCMs) have been applied to predict climate (Lim et al. 2009). Based on the observational relationships of dependent variable with various independents, the empirical models with some limitations are still the most widely used methods for monthly and seasonal rainfall predictions in agricultural planning because of the GCMs’ complexity (Meinke et al. 2007). The logic behind empirical methods is to identify relevant features of previous records of predictand and predictor variables and to apply these to predict values in the future. As one of the empirical approaches, artificial neural network (ANN) has been widely employed in water resources modeling (Abbot and Marohasy 2012; Venkata Ramana et al. 2013; Gholami et al. 2015; He et al. 2015). However, ANN also has its own shortcomings, such as slowly learning speed, convergence to local minimum, overfitting, and the components of the model’s complex structure (Okkan and Serbes 2012; Huang et al. 2014). The support vector machine (SVM) was proposed by Cortes and Vapnik (1995) on the basis of statistical learning theory and the structural risk minimization for objective functions that can theoretically achieve the global optimum (Chau and Wu 2010; Huang...
et al. 2014). In the last two decades, the SVM has been successfully applied in the field of hydrology for hydrologic predictions (Yu et al. 2006; Huang et al. 2014; Jajarmizadeh et al. 2015). Yu et al. (2006) employed the SVM to predict real-time flood stages in the Lan-Yang River, Taiwan. Huang et al. (2014) proposed a modified empirical mode decomposition (EMD)-based SVM model for monthly streamflow forecasting in the Wei River basin, China. Jajarmizadeh et al. (2015) used the SVM to predict the monthly streamflow in the Roodian watershed, Iran. Least squares support vector machines (LSSVMs), a modified version of SVM, convert the quadratic optimization problem into a system of linear equations, so that it has a computational advantage over the SVM when the number of predictor variables increases. The LSSVM can improve the forecast model’s anti-noise ability (Y. Wang et al. 2015b) and has been used in rainfall–runoff and streamflow prediction problems (Okkan and Serbes 2012; Kisi 2015). Okkan and Serbes (2012) used the LSSVM to predict the runoff values of the Tahtali and Gordes watersheds and found that the running time of LSSVM is significantly faster with the same or higher accuracy than that of ANN, autoregressive moving average (ARIMA), and multiple linear regression models. Kisi (2015) applied the LSSVM to estimate the monthly streamflow of the Dicle basin in Turkey and showed that the LSSVM performs better than adaptive neuro-fuzzy inference system–embedded fuzzy c-means clustering (ANFIS-FCM) and ARIMA. Therefore, it is very attractive to develop LSSVM-based models to forecast monthly rainfall.

The precipitation forecast models usually use a set of attribute values as inputs to predict precipitation. An input set of attribute values may comprise antecedent precipitation and/or other lagged climate-related values, which results in a large set of candidate inputs and complex computational process. For example, Hartmann et al. (2008) used a set of lagged climate signals, including the Southern Oscillation index; the Scandinavia, polar/Eurasia, and East Atlantic/western Russia patterns; and 11 indices calculated from snow data, sea surface temperatures (SSTs), and sea level pressures (SLPs) to forecast summer rainfall in the Yangtze River basin. Peng et al. (2014) used a set of lagged climate indices, including two large-scale atmospheric circulation modes [the Arctic Oscillation (AO) and North Atlantic Oscillation (NAO)] and six indices derived from the SSTs in the tropical central and eastern Pacific, South China Sea, Kuroshio, western Pacific warm pool, and Indian Ocean to forecast seasonal precipitation over China.

However, the performance of models is heavily dependent on the input variables used to develop the precipitation forecast models. In fact, it is essential to select a small and appropriate set of model inputs from available candidates for reducing the number of free parameters and improving computational efficiency of models. In recent years, a number of input selection algorithms have been proposed to identify significant input variables for model development (Sharma 2000; Sharma and Mehrotra 2014; Taormina and Chau 2015). For example, Sharma (2000) presented a partial mutual information (PMI) algorithm to select significant variables from a set of candidates, and then Sharma et al. (2000) used the algorithm in formulating predictors for hydrologic forecast. In the work of Bowden et al. (2005), it has been shown that the PMI method outperforms the common methods used to identify the model inputs. Recently, the PMI logic was extended to the partial information (PI) algorithm by Sharma and Mehrotra (2014) for determining an appropriate set of model inputs without assuming model representation. Based on the above, in this study we use the PI algorithm to identify the appropriate model inputs to LSSVM from a large set of available candidates.

The optimization of model parameters is a general question raised on designing SVM models. The resampling is generally used as a possible method for optimizing the SVM parameters by a trial and error procedure. However, the procedure is computationally expensive for tuning several SVM parameters. In recent years, a number of optimization methods, such as genetic algorithm (GA; Pai and Hong 2005), particle swarm optimization (PSO; Sudheer et al. 2014), and differential evolution (DE; Wang et al. 2012) have been applied to identify the optimal SVM parameters. Among these methods, the DE is one of the most efficient methods for optimizing the model parameters. The method is a stochastic search and optimization algorithm that uses a special differential operator to generate offspring vectors from a parent instead of the classical mutation operations (Storn and Price 1995). As a recent optimization technique, the DE algorithm has been successfully used in many parameter optimization problems (Maulik et al. 2010; Wang et al. 2012; Chen et al. 2015). Vesterstrom and Thomsen (2004) demonstrated that the DE algorithm performs better than the GA and PSO in the examination of the 34 widely used benchmark functions. Storn and Price (1997) also showed that the DE outperformed considerably both simulated annealing and the simplex method, and was equal or superior to the common evolutionary algorithms including GA, evolution strategies, and genetic programming. Chen et al. (2015) reported that the DE algorithm attained the best performance in generalization and forecasting and provided weak superiority over two other population-based
optimization algorithms, artificial bee colony (ABC) and ant colony optimization (ACO). In this study, the DE algorithm is adopted to determine the appropriate parameters in LSSVM models for improving the precipitation forecasting accuracy.

The precipitation generation process generally is highly nonstationary, nonlinear, and time varying so that the precipitation series contains different frequency components (X. G. He et al. 2014). In some previous studies (Guo et al. 2011; Okkan and Serbes 2012), the original series were directly used as the input variables when the forecasting models were developed, which leads to missing some features of different resolutions. To improve the results of the empirical models, the wavelet decomposition (WD; Schaeffli et al. 2007; Rivera et al. 2012) and empirical mode decomposition (Huang et al. 2014) are suggested to deal with the input and/or output data preprocessing in the time domain, and some successful results have been provided by previous studies in many areas including climate and hydrology (McMahon et al. 2008; Wu et al. 2009; Huang et al. 2014; Liu et al. 2015). Wu et al. (2009) introduced a hybrid model coupling ANN with the WD for the daily flows prediction. Liu et al. (2015) forecasted streamflow with rainfall and climate information inputs by developing a hybrid SVM model with the WD. McMahon et al. (2008) proposed an EMD-based approach to stochastically generating six-monthly rainfall sequences in Canberra, Australia. Karthikeyan and Kumar (2013) assessed the predictability of nonstationary time series using WD- and EMD-based autoregressive moving average (ARMA) models in forecasting hydrologic data. Huang et al. (2014) proposed a modified EMD–SVM model for monthly streamflow prediction in the Wei River basin and showed that the modified model has a good stability and a high prediction precision. These previous studies indicated that the hybrid models, which were trained and validated with the decomposed data by the WD or EMD, provided better results than the reference models without decomposed series. In practice, the application of wavelets demands the choice of basis functions, requiring considerable skill in their selection and application (Huang et al. 2014). However, the implementation of EMD does not need any predetermined basis function, and the main advantage of EMD over traditional approaches is its complete self-adaptiveness and its very local ability both in physical space and frequency space (W. C. Wang et al. 2015), so the EMD is chosen for decomposing nonlinear and nonstationary data in this study.

In this study, we develop a new hybrid least squares support vector machine (HLSSVM) model, which is an effective integration of four different methods (EMD, PI, DE, and LSSVM), for predicting monthly precipitation from lagged precipitation and climate indices. The hybrid method is designed by using the EMD to decompose the predictand and predictor variables into a set of component subseries, the PI algorithm to select a small and appropriate subset of inputs from a set of candidate predictor subseries, and the DE to determine the optimal parameters of the LSSVM. The proposed hybrid method is tested and compared to the LSSVM and LSSVM–DE models using 138 meteorological stations in the Yangtze River basin. We expect that the coupled HLSSVM can inherit the advantages from the LSSVM, DE, and EMD and improve the monthly precipitation forecasting accuracy compared to the reference models.

The rest of this paper is organized as follows. In section 2, we describe the details of the study area and data. In section 3, a brief introduction to the methods mentioned above, including the EMD, LSSVM, PI, and DE, is provided and a new hybrid least squares support vector machine is presented. In section 4, the proposed hybrid method is examined by forecasting monthly precipitation at 138 meteorological stations across the Yangtze River basin and compared with the LSSVM and LSSVM–DE models. Some conclusions are made for this research in section 5.

2. Study area and data collection

a. Study area

To evaluate the forecasting performance of the proposed model, the HLSSVM is used to forecast monthly precipitation at 138 meteorological stations over the Yangtze River basin, as shown in Fig. 1. The Yangtze River basin (24°27′–35°54′N, 90°33′–122°19′E), which has a diverse range of climate zones, is the largest in China and the third largest in the world. It is well known that the southern part of the basin is climatically close to subtropical climate and northern part to the temperate zone. The annual average precipitation in the whole basin is about 1100 mm, varying from 270 to 500 mm in the western region of the basin and 1600 to 1900 mm in the southeastern region (Gemmer et al. 2008). The Yangtze River basin experiences annually a cold, dry winter and a warm, wet summer with 70%–80% of its annual rainfall (Yu et al. 2009). Precipitation declines across the catchment from east to west, and the upper Yangtze region with highest elevation has the least precipitation (Chen et al. 2014). Referring to Hartmann et al. (2008) and Peng et al. (2014), we select the previous precipitation and 11 large-scale climate indices as a set of potential input variables to forecast monthly precipitation over the Yangtze River basin.
b. Precipitation data

Long-term monthly precipitation data were downloaded from the China Meteorological Data Sharing Service System (http://data.cma.cn/). There are 138 meteorological stations used in this study, with data covering from January 1960 to December 2012 distributed relatively uniformly across the Yangtze River basin (Fig. 1). All monthly precipitation time series are transformed to standardized monthly precipitation anomaly values for prediction. The climatological mean and standard deviation are first computed for each of the 12 months over the 1960–2002 period for each precipitation series. Then, the standardized monthly precipitation anomaly series can be derived by subtracting the climatological mean of the corresponding month from each observation and dividing by the climatological standard deviation of the month.

c. Climate indices

In this study, 11 climate indices, which have been identified to have teleconnections with the variability of precipitation over China, are chosen as potential predictor variables for forecasting monthly precipitation in the Yangtze River basin. A brief description of the 11 climate indices is given in Table 1. They include HadISST Niño-3 (Niño-3), HadISST Niño-3.4 (Niño-3.4), Oceanic Niño Index (ONI), average SST anomaly over the Bay of Bengal (BB), average SST anomaly over the South China Sea (SCS), average SST anomaly over the East China Sea (ECS), average SST anomaly over the Kuroshio (KC), average SST anomaly over the western Pacific warm pool (WPWP), Indian Ocean dipole mode index (DMI), AO, and NAO. The SST is usually identified as the most important and reliable predictor for seasonal climate forecasts owing to its long-term persistence with respect to atmospheric processes (Lang and Wang 2010). For example, the SST anomalies in the tropical central and eastern Pacific Ocean known as El Niño–Southern Oscillation (ENSO) variables have impacts on the seasonal variability of precipitation over China. Xiao et al. (2015) found that ENSO is the leading driver of seasonal precipitation variability in the Yangtze River basin. Three different ENSO indices, ONI, Niño-3, and Niño-3.4, are selected as potential predictor variables in this study. Besides the SSTs related to ENSO, the SSTs over other regions, such as the Bay of Bengal, South China Sea, East China Sea, Kuroshio, western Pacific warm pool, and Indian Ocean, have also been revealed to have strong influences on precipitation over China (Hartmann et al. 2008; Peng et al. 2014). Hartmann et al. (2008) employed the SSTs over the BB, SCS, and ECS to forecast summer rainfall in the Yangtze River basin. Peng et al. (2014) revealed that the SSTs over the South China Sea, Kuroshio, western Pacific warm pool, and the Indian Ocean have strong impacts on the seasonal variability of precipitation over China. In addition, the large-scale atmospheric circulation modes, such as AO and NAO, also have an influence on precipitation over China both individually or in conjunction with SSTs. Xiao et al. (2015) found that the winter precipitation in the Yangtze River basin is influenced by the NAO. Gong and Ho (2003) indicated the connection between the AO and the variability of year-to-year East Asian summer monsoon in parts of China. Therefore, the AO and NAO are also selected as potential predictors of precipitation anomaly in the Yangtze River basin in this study.

3. Methodology

a. Information theory

In this subsection, we introduce the information theory metrics including the information entropy, mutual information, and partial information. The information entropy is used to measure the statistical uncertainty and predictability of monthly precipitation series, and the
mutual information and partial information are applied to identify the most significant predictors for monthly precipitation forecasts from a set of candidate predictors.

The entropy proposed by Shannon (1948) has been used in a wide range of applications assessing variability in the hydrological variables and has significant potential for quantifying the inherent predictability of the hydrological cycle (DeSole and Tippett 2007). The entropy of a random variable $X$ is defined as

$$ H(X) = -\int f(x) \log[f(x)] \, dx, \tag{1} $$

where $f(x)$ is the probability density function (PDF) of $X$. In general, higher entropy indicates higher statistical uncertainty and lower predictability of time series and vice versa (Elsner and Tsonis 1993; Li and Zhang 2008).

The mutual information (MI) is a measure of statistical dependence between variables and is able to detect any type of functional relationship. The MI between $Y$ and $X$ is defined as (Fraser and Swinney 1986)

$$ \text{MI}(Y, X) = \int f_{Y,X}(y, x) \log \left[ \frac{f_{Y,X}(y, x)}{f_Y(y) f_X(x)} \right] \, dy \, dx, \tag{2} $$

where $f_Y(y)$ and $f_X(x)$ are the marginal PDF of $Y$ and $X$, respectively, and $f_{Y,X}(y, x)$ is the joint PDF of $Y$ and $X$. The MI defined in (2) can be easily extended to measure the dependence for multivariate. However, it cannot be used to compute the partial dependence among multiple variables. Thus, the limited use of MI to measure the partial dependence conditional to preidentified predictors precludes its use in model input identification. Such identification usually starts from a null model without predictors and involves adding variables such that the predictive uncertainty is reduced. To address this problem, Sharma and Mehrotra (2014) proposed a stepwise selection algorithm for system predictor identification by developing the partial information measure.

The PI, which represents the partial dependence of $Y$ on $X$ conditional to the preselected predictor variable set $Z$, is given by

$$ \text{PI}(Y, X \mid Z) = \int [f_{Y,Z}(y, z)] \log \left[ \frac{f_{Y,Z}(y, z)}{f_Y(y) f_Z(z)} \right] \, dy \, dz, \tag{3} $$

where $z$ is a realization of $Z$. The PI in (3) measures the conditional dependence between $Y$ and $X$ after considering the effect of a preselected predictor variable or vector $Z$. By representing the above conditional PDFs as ratios of the respective joint and marginal PDFs, a sample estimate of the PI in (3) $[\hat{\text{PI}}(\cdot)]$ can be obtained by (Sharma and Mehrotra 2014)

$$ \hat{\text{PI}}(Y, X \mid Z) = \hat{\text{MI}}(Y, X, Z) - \hat{\text{MI}}(Y, Z) \tag{4} $$

where $\hat{\text{MI}}(\cdot)$ is the sample estimate of MI(·). For additional details on the estimation of the MI, readers are referred to Sharma (2000) and Sharma and Mehrotra (2014).

### b. Input identification algorithm

Following the main ideas in Sharma and Mehrotra (2014), we sketch the PI stepwise selection algorithm in this subsection. The algorithm starts with an empty
preselected input vector \( \mathbf{Z} \), along with a set of candidate predictor variables \( \mathbf{X} \). The model input identification proceeds according to the basic steps below:

1) Estimate the sample PI correlation (PIC) between output \( Y \) and each input variable in \( \mathbf{X} \) by

\[ \widehat{\text{PIC}} = \sqrt{1 - \exp(-2\widehat{\pi})} \]

Denote the highest of these estimates as \( \text{PIC}^\ast \) and the associated variable as \( X^\ast \).

2) Estimate the threshold \( \text{PIC}_\alpha \) for a given significance level \( \alpha \) of the statistics. Here, the \( \text{PIC}_\alpha \) is derived by assuming \( T = \text{PIC} \sqrt{m/(1 - \text{PIC}^2)} \) that follows the Student’s \( T \) distribution with \( m = n - l \) degrees of freedom, where \( n \) is the total number of samples and \( l \) the number of conditioning variables. Then, the \( \text{PIC}_\alpha \) is used to judge whether the \( \text{PIC}^\ast \) in step 1 is significant or not.

3) If \( \text{PIC}^\ast > \text{PIC}_\alpha \), the corresponding \( X^\ast \) is a significant predictor and is added to preexisting predictor set \( \mathbf{Z} \), and remove \( X^\ast \) from \( \mathbf{X} \).

4) Repeat steps 1–3 as many times as needed.

5) The algorithm stops once all significant predictors have been included in \( \mathbf{Z} \).

c. Least squares support vector machine

The LSSVM is briefly described as follows. More detailed information about the LSSVM is available from Suykens and Vandewalle (1999) and Suykens et al. (2002). Given a set (denoted by curly brackets) of training data \( \{(x_j, y_j)\}_{j=1}^J \), where \( x_j \in \mathbb{R}^n \) is an input vector with a dimension of \( n \), \( y_j \in \mathbb{R}^1 \) is a corresponding target output and \( J \) represents the size of the training data. The regression function that relates the input vector to the output can be formulated as (Suykens et al. 2002)

\[ y = f(x) = \omega^T \varphi(x) + b, \]

where \( \varphi(x) \) expresses a nonlinear transfer function mapping from the input vector to a high-dimensional feature space, \( \omega \) is a weight vector of the hyper plane, and \( b \) means a bias. Transforming the regression problem (5) into a constrained quadratic optimization problem, \( \omega \) and \( b \) can be estimated by minimizing the cost function. The cost function of the LSSVM is defined as

\[ \min \left( \frac{1}{2} \omega^T \omega + \frac{c}{2} \sum_{j=1}^J \xi_j^2 \right) \]

subjected to the following constraints:

\[ y_j - [\omega^T \varphi(x_j) + b] = \xi_j, \quad j = 1, 2, ..., J, \]

where \( c \) is the regularization parameter and \( \xi_j \) represents the error for \( x_j \). The solution of the optimization problem of LSSVM is obtained by regarding the Lagrangian as

\[ L(\omega, b, \xi, \alpha) = \frac{1}{2} \omega^T \omega + \frac{c}{2} \sum_{j=1}^J \xi_j^2 - \sum_{j=1}^J \alpha_j [\omega^T \varphi(x_j) + b + \xi_j - y_j], \]

where \( \alpha_i \) is a Lagrange multiplier. The conditions for optimality can be obtained by taking the partial derivatives of (8) with respect to \( \omega \), \( b \), \( \xi \), and \( \alpha \), respectively. By eliminating \( \omega \) and \( \xi \), the solution of optimization problem can be written as

\[ \begin{pmatrix} 0 \\ I^T \\ \Omega + c^{-1} I \end{pmatrix} \begin{pmatrix} b \\ \omega \\ \alpha \end{pmatrix} = \begin{pmatrix} y \\ \Phi \end{pmatrix}, \]

where \( \alpha = (\alpha_1, \alpha_2, ..., \alpha_J)^T \), \( y = (y_1, y_2, ..., y_J)^T \), \( \Phi = \varphi(x_1) \varphi(x_J) \). Consequently, the resulting LSSVM for function estimation can be expressed as

\[ \hat{y} = \sum_{j=1}^J \alpha_j^* \varphi(x, x_j) + b^*, \]

where \( \varphi(x, x_j) = \varphi(x) \varphi(x_j) \) is the kernel function, and \( \alpha^*_j \) and \( b^* \) are the solutions to (9).

There are several general kernel functions that are being used in LSSVM, such as linear kernel, polynomial kernel, radial basis function (RBF) kernel, and sigmoid kernel. Compared with these kernel functions, the RBF kernel is the most commonly used kernel function in SVM model, which is mapping the sample set from the input space into a high-dimensional feature space with high efficiency (Wang et al. 2012; Z. B. He et al. 2014; Zhang et al. 2015). The RBF kernel can be written as

\[ K(x, x_j) = \exp(-\gamma \|x - x_j\|^2), \]

where \( \gamma > 0 \) denotes the kernel width. In this paper, the RBF kernel is selected as the kernel function of the LSSVM because of its simple structure and strong generalization ability after using different kernels for the models. Regularization parameter (i.e., \( c \)) and the kernel width (i.e., \( \gamma \)) should be set properly for models, because they can prevent the models from overfitting the data with efficient performance (Zhang et al. 2015). An automatically efficient method, differential evolution algorithm, is used to determine the appropriate parameters in LSSVM model in this study.
d. Differential evolution algorithm

The DE is one of the evolutionary algorithms, which was developed by Storn and Price (1995) to solve the Chebyshev polynomial fitting problem. Now, the method has become a popular optimization algorithm and has been successfully applied in the parameter optimization of the SVM (Wang et al. 2012). A flow diagram of the entire process of DE is shown in Fig. 2 for optimizing the parameters of LSSVM. According to Storn and Price (1997) and Mayer et al. (2005), the brief introduction of DE is given as follows.

First, set the upper and lower bounds of the parameter vector \((c, \gamma)\) of the LSSVM and the population size \(N\) and maximum generation number \(G\) of the DE. Then, the initial population \(V_{i,0} = (c_{i,0}, \gamma_{i,0}), i = 1, 2, \ldots, N,\) is generated randomly within the bounds and should cover the entire parameter space. Next, run the LSSVM model with the generated population to get the rainfall forecasting, and then from the forecast results, compute and record the fitness function values of all the individuals. For each of the population members in turn, the offspring is generated. If the offspring has superior fitness values, it will replace the parent vector in the next generation. Otherwise, the parent survives and is passed on to the next generation of the algorithm. Therefore, a new generation with better mean fitness is constituted. The process is continued in this way to achieve convergence close to an optimal solution. In the DE algorithm, the offspring \(V_{ig+1}, g = 0, 1, \ldots, G - 1,\) is generated by the three steps including mutation operation, crossover operation, and selection operation. More specifically, DE’s basic strategy can be obtained from Storn and Price (1997).

e. Empirical mode decomposition

The EMD has become the adaptive and valuable tool in nonlinear and nonstationary time series decomposition and has been widely applied in the field of hydrology because of its simplicity and lesser computational cost (Karthikeyan and Kumar 2013). The EMD is derived from the simple assumption that any time series consists of a residual component and a finite number of independent intrinsic mode functions (IMFs). Each IMF represents an embedded characteristic oscillation of the original signal on a separated time scale. An IMF must satisfy two conditions (Huang et al. 1998): 1) the number of extreme values and that of zero crossings must be equal or at most differ by one in the whole signal segment, and 2) at any point, the mean of envelope defined by the local minima and that of the local maxima must be zero. Using EMD, a time series \(x(t)\) can be decomposed into some IMFs \(s_i(t), i = 1, \ldots, I,\) where \(I\) is the highest decomposition level, and a residual is \(r_I(t)\). Then, the original series \(x(t)\) can be reconstructed from the IMFs and residual as

\[
x(t) = \sum_{i=1}^{I} s_i(t) + r_I(t).
\]

The EMD implementation of the decomposition of time series into IMFs is an iterative procedure called sifting, which is used to remove riding waves and make the wave profiles more symmetric. The basic steps to successively extract the IMFs are as follows (Yang et al. 2010; Huang et al. 2014):

1) Extract all the local extreme values including local minima and maxima of \(x(t)\).
2) Connect all local minima and maxima to obtain the upper and lower envelopes \(e_{up}\) and \(e_{low}\), respectively, by means of cubic splines or more refined methods.
3) Calculate the mean \(m(t)\) of the upper and lower envelopes by \(m(t) = (e_{up} + e_{low})/2\).
4) Define the detail \( h(t) = x(t) - m(t) \), and check whether \( h(t) \) is an IMF or not. If \( h(t) \) meets the above two conditions, then an IMF \( h(t) \) is obtained with the residual \( r(t) = x(t) - h(t) \) replacing \( x(t) \), else \( h(t) \) is not an IMF with \( h(t) \) replacing \( x(t) \).

5) Repeat steps 1–4 until a given stopping criterion is satisfied, and set \( s_1(t) = h(t) \), which is the first IMF. More detailed information about the stopping criterion can be obtained from Huang et al. (1998) and Rilling et al. (2003). In this study, the stopping criterion recommended by Rilling et al. (2003) is used.

6) Repeat steps 1–5 until all the IMFs are found.

Figure 3 shows an example of EMD with five resolution levels corresponding to the original series for monthly precipitation anomaly of station 56144 (31.80°N, 98.58°E) and Niño-3 index. This figure clearly shows how the original series is decomposed into its intrinsic mode functions and residual subseries. From Fig. 3, we can observe that the IMFs delineate varying frequencies, amplitudes, and wavelengths. The first IMF exhibits the highest frequency, maximum amplitude, and shortest wavelength. The following IMF components show successively decreasing frequencies and amplitudes and increasing wavelengths. The residue component is a mode slowly changing around the long-term average.

f. Hybrid least squares support vector machine

In this subsection, a new HLSSVM model is proposed by coupling four different methods (EMD, PI, DE, and LSSVM) and is used to forecast monthly precipitation. The architecture of the HLSSVM model is illustrated in Fig. 4. The entire procedure for the HLSSVM forecasting consists of four steps, namely, EMD, input variable identification, parameter optimization, and SVM modeling. Each of the predictor and predictand time series is first decomposed by EMD into a residual component \( r \) and a certain number of independent orthogonal components \( s_i \) at different decomposition levels. Then, at each level, an appropriate subset of model inputs is identified by the PI algorithm from a large set of potential predictor subseries for forecasting predictand subseries at the corresponding level. In the third step, the parameters of LSSVM models are optimized by DE algorithm. Finally, the optimized models at the different decomposition levels are separately employed to determine the relationship between the predictand and identified predictor subseries in the training period and then forecast the future predictand subseries.

Specifically, for forecasting monthly rainfall by the HLSSVM, the estimated \( \hat{r}_i^y(t_k) \) of the residual of the rainfall anomaly can be computed by

\[
\hat{r}_i^y(t_k) = f[r_i^{x_i}(t_{k-s_{i,3}}), r_i^{x_i}(t_{k-s_{i,2}}), \ldots].
\]
where the function $f$ is determined by Eq. (10) from the training period; $r^x_{ij}$ represents the residual of the predictor $x_i$, which is selected from a set (denoted by curly brackets) of candidates $\{r^x_{ij}\}_{j=1}^J$ by PI algorithm; $s_{ij} \geq 1$ is the time lag between the precipitation anomaly residual $r^x_{ij}$ and predictor residual subseries $r^x_{ij}$; and $J$ is the highest decomposition level. Similarly, the estimated $\hat{y}_i(t_k)$, $i = 1, \ldots, I$, is the time lag between the precipitation anomaly residual $r^y_{I}$ and predictor residual subseries $r^x_{I}$. By PI algorithm.

$$\hat{y}_i(t_k) = f[s^x_i(t_{k-\tau}), s^x_i(t_{k-\tau}), \ldots],$$

where $s^y_i$ stands for the independent component of the predictor $x_i$ at the decomposition level $i$, which is selected from a set (denoted by curly brackets) of potential predictors $\{s^x_i\}_{j=1}^J$ and $\tau_{ij} \geq 1$ is the lag time. According to (12)–(14), the forecasted $\hat{y}(t_k)$ of the precipitation anomaly can be expressed as

$$\hat{y}(t_k) = \sum_{i=1}^I \hat{s}^y_i(t_k).$$

We would like to note that to predict each precipitation anomaly subseries, an LSSVM model is constructed in which the model parameters ($c$ and $\gamma$) are optimized by a DE algorithm in the training period. Finally, the rainfall forecasts can be calculated from the sum of the forecasted rainfall anomaly subseries by the inverse transform of standardized monthly anomaly.

g. Implementation and assessment of models

MATLAB codes were written for the HLSSVM simulation based on the LS-SVMlab toolbox (De Brabanter et al. 2011). Then, the forecasting performance of the HLSSVM model is evaluated by two statistical indices, the Nash–Sutcliffe efficiency (NSE) and relative absolute error (RAE), and compared to that of LSSVM and LSSVM–DE models. The reference models use the undecomposed time series as potential inputs, and the significant predictors are selected by the PI algorithm. Here, the reference LSSVM is implemented with default options in the LS-SVMlab toolbox. The reference LSSVM–DE is formed by using DE algorithm to optimize the parameters of LSSVM model. The prediction performance of all the models is measured by NSE and RAE, which are respectively defined as

$$\text{NSE} = 1 - \frac{\sum_{i=1}^M (y_{\text{for},i} - y_{\text{obs},i})^2}{\sum_{i=1}^M (y_{\text{obs},i} - \bar{y}_{\text{obs}})^2}$$

and

$$\text{RAE} = \frac{\sum_{i=1}^M |y_{\text{for},i} - y_{\text{obs},i}|}{\sum_{i=1}^M y_{\text{obs},i}}.$$
anomaly and large-scale climate index series at lags of 1–12 months are used as potential model inputs in this study.

4. Results and discussion

There are a total of 144 potential initial inputs for each of the SVM models in the HLSSVM by using the predictors with lags of 1–12 months. The PI selection algorithm with significance level $\alpha = 0.01$ has identified 8–9 input variables for each SVM as significant predictors for each of the 138 rainfall stations over the Yangtze River basin. Some useful predictor variables may have been excluded by using the rule with $\alpha = 0.01$. To examine whether or not more predictors in the final models help to improve the model skill, the PI algorithm with a looser criterion $\alpha = 0.05$ is performed, which results in more variables being chosen into significant predictor set. However, the prediction performance of models is not improved from the RAE and NSE viewpoints. Therefore, here we only provide the performance statistics of models with $\alpha = 0.01$. In addition, we note that all the models occasionally produce negative forecasting values for some extremely dry months. We adjust the negative values to zero rainfall as a physical constraint on the system before the RAE and NSE are computed.

The spatial distribution of the NSE and RAE are mapped in Figs. 5 and 6, respectively, for the HLSSVM, LSSVM–DE, and LSSVM models with the test period 2003–12 for 138 selected stations over the Yangtze River basin. It can be seen from Figs. 5 and 6 that the forecasting results obtained by these three models tell similar spatial patterns of the model forecast skill with a clear west and northwest–east and southeast distribution. As can be observed, the stations with larger NSE and smaller RAE values are mainly located in the western and northwestern regions of the basin, and eastern and southeastern regions are mainly covered by the smaller NSE and larger RAE values (Figs. 5, 6). Thus, it can be concluded that all the models perform better in the western and northwestern regions than in the eastern and southeastern regions of the Yangtze River basin. For example, the HLSSVM model acquires larger NSE values ($\geq 0.60$) at 49 out of 138 stations, 70% of which distribute in the western and northwestern regions of the basin. At the same time, the smaller NSE values ($< 0.60$) from the HLSSVM is identified at 89 stations, 90% of which locate in the eastern and southeastern regions of the Yangtze River basin (Fig. 5a). From Figs. 5d and 6d, we can also see that the entire basin is dominated by increased NSE and reduced RAE from the HLSSVM model relative to the LSSVM model. We note that the percentages of increased NSE and decreased RAE against those of the LSSVM range from 0.3% to 112.6% and from 0.6% to 21.4%, respectively, at more than 90% of stations.

To assess the relationship between randomness and predictability of the precipitation series in the Yangtze River basin, we computed the entropy value of monthly precipitation series for each station. The spatial distribution of the entropy for 138 selected stations across the Yangtze River basin is mapped in Fig. 7. It can be seen that Fig. 7 exhibits a similar spatial distribution with Figs. 5 and 6. The entropy values in the western and northwestern regions of the basin are lower than those in other regions and the higher values mainly occur in the western and southeastern regions. This indicates that the western and northwestern areas have lower randomness of monthly precipitation and a less degree of complexity for precipitation prediction. Comparing Figs. 5 and 6 with Fig. 7, we can observe that the spatial pattern of model skill is significantly correlated with the spatial distribution of information entropy of precipitation series. The forecast of the models is more skillful at the stations located in the western and northwestern regions with smaller entropy values, and the forecast skill is limited at the stations in the eastern and southeastern regions with higher entropy values. However, from Figs. 5d and 6d, we can also see that relative to the LSSVM model, the HLSSVM provides a more significant improvement for monthly rainfall forecasts at most of the stations in the eastern and southeastern regions with higher entropy values. This indicates that the forecast skill of the HLSSVM is more robust than that of the LSSVM.

To further reveal the relationship between the predictability and the randomness of the rainfall series, Fig. 8 presents the performance statistics of models as a function of the entropy, which shows a strong inverse correlation between the NSE and the entropy (Fig. 8a) and a strong positive correlation between the RAE and the entropy (Fig. 8b). Here, we would like to emphasize that lower entropy indicates lower randomness and higher predictability of time series, and larger NSE and smaller RAE imply higher forecast performance of the models. From Fig. 8, we can observe that the lower the entropy of precipitation series is, the better the forecast performance of the models is. The correlation coefficient $R$ between the NSE and the entropy are $-0.464$, $-0.514$, and $-0.511$ for HLSSVM, LSSVM–DE, and LSSVM, respectively (Fig. 8a). The correlation between the RAE and the entropy for HLSSVM is $R = 0.325$, for LSSVM–DE is $R = 0.418$, and
and for LSSVM is $R = 0.423$ (Fig. 8b). It should be noted that all these correlation coefficients are statistically significant at the significance level $\alpha = 0.01$. The comparison of the correlation coefficients reveals that the relationship between the entropy and the performance statistics of HLSSVM is weaker than those for the LSSVM–DE and LSSVM. This is because the HLSSVM gives the improved forecasts at stations with higher entropy values relative to two reference models (Fig. 8). Therefore, it may be inferred that compared with the LSSVM–DE and LSSVM, the HLSSVM has more capability to capture the useful information contained in the time series with higher randomness. The slopes of the linear trend lines from the different models also indicate that the LSSVM–DE has a better stability than the LSSVM, and the stability of the HLSSVM is best for forecasting monthly precipitation with different entropy values.

Figure 9 depicts the frequency histograms of the number of stations for the NSE and RAE obtained by...
three different models over the basin. From Fig. 9, we can find that the HLSSVM model performs significantly better than the LSSVM–DE and LSSVM, and the performance of the LSSVM is poorest among these three models. More specifically, the Nash–Sutcliffe efficiency by the HLSSVM ranges from 0.27 to 0.83 (Fig. 5a), with approximately 98% of stations acquiring an NSE $\geq 0.30$ and 60% of stations acquiring an NSE $\geq 0.50$ (Fig. 9a), and the NSE by the LSSVM–DE varies from 0.21 to 0.84 (Fig. 5b), acquiring about 95% of stations with an NSE $\geq 0.30$ and 51% of stations with an NSE $\geq 0.50$ (Fig. 9a). However, the NSE by the LSSVM ranges from 0.14 to 0.82 (Fig. 5c), with about 82% stations obtaining an NSE $\geq 0.30$ and only 43% of stations obtaining an NSE $\geq 0.50$ (Fig. 9a). Therefore, the HLSSVM increases the predicting NSE by 0.01–0.25 at 91% of stations and by 0.11–0.25 at 31% of stations relative to LSSVM (Fig. 5d). The average NSE values calculated from all stations are 0.48, 0.52, and 0.55 for the LSSVM, LSSVM–DE, and HLSSVM, respectively. These clearly
indicate that the LSSVM–DE performs better than the LSSVM, and the HLSSVM is the most effective model and provides improved prediction accuracy compared with the LSSVM and LSSVM–DE. The relative absolute error by the HLSSVM model varies between 0.26 and 0.52 (Fig. 6a), with an RAE $\leq 0.50$ being identified at only 9% of the stations and an RAE $\leq 0.42$ at about 58% of the stations (Fig. 9b). The RAE by the LSSVM–DE model ranges from 0.27 to 0.55 (Fig. 6b), acquiring about 19% of stations with an RAE $\geq 0.50$ and 42% of stations with an RAE $\leq 0.42$ (Fig. 9b). However, the RAE by the LSSVM model varies from 0.27 to 0.58 (Fig. 6c), with an RAE $\geq 0.50$ being found in approximately 30% of the stations and an RAE $\leq 0.42$ at only 36% of stations (Fig. 9b). Thus, the HLSSVM reduces the predicting RAE by 0.01–0.11 at 93% of stations and by 0.05–0.11 at 30% of stations in comparison to the LSSVM (Fig. 6d). The average RAE values over the basin are 0.44, 0.42, and 0.40 for the LSSVM, LSSVM–DE, and HLSSVM, respectively. Therefore, from the NSE and RAE statistics, the proposed HLSSVM model outperforms the LSSVM and LSSVM–DE, and the LSSVM has the poorest performance among three models for monthly precipitation forecasts in the Yangtze River basin. The LSSVM–DE model uses the DE algorithm to optimize the parameters of LSSVM, so it gets a better performance than the LSSVM. In the HLSSVM model, both the original predictor and predictand series are decomposed by EMD into subseries under the different scales. As pointed out by Huang et al. (2014), each of the decomposed subseries plays a different role in the original series and the feature of each subseries is obvious. Thus, the hybrid model can optimally utilize the information contained in the different subseries and is able to capture the influences of the predictors on the precipitation anomaly under the different time scales. This may be why the HLSSVM model performs better than the LSSVM and LSSVM–DE.

To examine how the detailed monthly forecasts match with the observations, Fig. 10 depicts the forecasted results obtained by the different models in the test period 2003–12 for two stations highlighted by red dots in Fig. 1: 56172 (31.90°N, 102.23°E) in the western region and 57662 (29.05°N, 111.68°E) in the eastern region of the basin. As shown in Fig. 10, all the models provide a reasonable monthly rainfall forecast at these two stations. At the same time, we also observe that the forecast by LSSVM–DE is closer to the observed value than that by the LSSVM, and the forecast by HLSSVM is most matched with the observed value among all the models. Although these three models provide similar forecasts for some months with dry or normal precipitation in Fig. 10, the HLSSVM model performs better compared with the LSSVM and LSSVM–DE for some extreme wet months [e.g., 2012 (Fig. 10a) and 2005–07 (Fig. 10b)]. We note that dry and wet conditions in each month are defined independently for each station on the basis of standardized monthly precipitation anomaly at that station. This superior performance of HLSSVM is clearer at the eastern station 57662 with higher entropy (Fig. 10b). For the first peak precipitation in 2012 at western station 56172, the forecasted 121.0 mm by the LSSVM has an underestimation of 85.5% relative to the measured 224.4 mm, and the forecasted 128.6 mm by LSSVM–DE has an underestimation of 74.5%, whereas the forecasted 171.2 mm by the HLSSVM only has an underestimation of 31.1% (Fig. 10a). Obviously, for some peak precipitation at eastern station 57662, the HLSSVM also gives improved forecasts compared with the LSSVM–DE and LSSVM (Fig. 10b). The scatter-plots of predictions by three models at western station 56172 and eastern station 57662 in the test period 2003–12 are also presented in Fig. 11. As we can see from Fig. 11, the forecasted values by the HLSSVM are closest to the 45° line compared with the LSSVM and LSSVM–DE. The improvement of the hybrid models is also demonstrated by the NSE values shown in Fig. 11.

To sum up, all the above model performance evaluations, including common measures of NSE and RAE at all selected stations (Figs. 5, 6, 8, and 9) and detailed
monthly precipitation predictions (Fig. 10) and scatter-plots of predictions at two selected stations (Fig. 11), indicate that the LSSVM–DE has a superior performance to the LSSVM and that the HLSSVM has the best accuracy among all the models in forecasting monthly precipitation over the basin.

5. Conclusions

In this paper, a hybrid least squares support vector machine model has been developed, examined, and discussed for effectively predicting monthly precipitation. The hybrid model is built by incorporating the empirical mode decomposition for preprocessing the input and output variables, partial information selection algorithm for identifying an appropriate set of model inputs, and differential evolution for optimizing model parameters into the least squares support vector machine models. The proposed model is evaluated by forecasting the monthly precipitation over the Yangtze River basin with the 53-yr calibration data from 1960 to 2002 and the 10-yr test data from 2003 to 2012 and is
compared to the LSSVM and LSSVM–DE models without data decomposition. We first observed that the forecast skill of all the models exhibits the similar spatial patterns in the Yangtze River basin. The models perform better in the western and northwestern regions than in other regions of the basin, which is strongly correlated with the spatial distribution of the randomness of precipitation series. In general, the lower the randomness of series is, the more skillful the forecast of the model is. Then, from the comparison of results, we concluded that the LSSVM model has the poorest performance, with the lowest NSE mean and highest RAE mean among three models. The LSSVM–DE has a better performance and it outperforms the LSSVM. Among all the models, the
HLSSVM has the best performance in terms of the two statistical measures, obtaining the highest NSE mean and lowest RAE mean, which increases the average NSE by 0.07 (with 0.11–0.25 increments at 31% stations) and reduces the average RAE by 0.04 (with 0.05–0.11 decrements at 30% stations) compared with LSSVM. The improvement is more significant at most of the stations located in the eastern and southeastern regions with higher randomness of precipitation. Thus, with the significant improvement and good robustness of monthly rainfall forecasts, the proposed hybrid model is an efficient tool for providing the reliable precipitation forecasts in agricultural planning, water resource management, and associated crop insurance application. However, it should be noted that the exercises provided in the study are monthly precipitation prediction with 1-month lead time in the Yangtze River basin. Therefore, it would be worthwhile to discuss how the HLSSVM performs for monthly precipitation forecasts with a longer lead time and/or for other regions of the world.

Although the proposed hybrid model in this study gives an encouraging perspective in monthly precipitation prediction, there exist some limitations and room for improving its potential capability. For some precipitation series with high randomness and complexity, the HLSSVM model may not substantially reduce the error between the forecasted and observed values, and the forecasted results may not be so accurate. To further enhance the robustness of hybrid model, we can incorporate other feasible time-decomposition methods and parameter optimization algorithms into the LSSVM (or other empirical models) to develop a more powerful tool for precipitation prediction. Further investigation of novel hybrid methods is worth pursuing in future.

Acknowledgments. We thank the anonymous reviewers for their careful reading and constructive comments that improved the paper. Funding for this study was provided by the National Natural Science Foundation of China Grants 41472238 and 41272271, Construct Program of the Key Discipline in Hunan Province of China Grant 20110001, Scientific Research Fund of Hunan Provincial Education Department of China Grant 14A097, and Aid Program for Science and Technology Innovative Research Team in Higher Educational Institutions of Hunan Province of China Grant 20130001. (The data for the model used in this study are freely available at http://www.cpc.ncep.noaa.gov and http://www.esrl.noaa.gov/psd/.)

REFERENCES


Liu, Z. Y., P. Zhou, and Y. Q. Zhang, 2015: A probabilistic wavelet-
——, ——, and J. X. Qin, 2015: A hybrid wavelet neural network
January 2017 T AO E T A L . 175
Lim, E. P., H. H. Hendon, D. Hudson, G. M. Wang, and O. Alves,
Huang, N. E., Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng,
Kisi, O., 2015: Streamflow forecasting and estimation using least
Taormina, R., and K. W. Chau, 2015: Data-driven input variable selection for rainfall–runoff modeling using binary-coded particle


