Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com

Research papers

# A hybrid support vector regression framework for streamflow forecast

Xiangang Luo[a], Xiaohui Yuan[b], Shuang Zhu[a], Zhanya Xu[a], Lingsheng Meng[c], Jing Peng[a]

[a] Faculty of Information Engineering, China University of Geosciences, Wuhan, China
[b] Department of Computer Science and Engineering, University of North Texas, Denton, TX 76210, USA
[c] Hubei Jinlang Survey and Design Co. LTD, Luoshi Road, Wuhan, Hubei 430074, China



ARTICLE INFO

This manuscript was handled by Emmanouil Anagnostou, Editor-in-Chief, with the assistance of Yiwen Mei, Associate Editor

*Keywords:*
Streamflow forecast
Regression
Factor analysis
Time series decomposition

ABSTRACT

Monthly streamflow time series are highly non-linear. How to improve forecast accuracy is a great challenge in hydrological studies. A lot of research has been conducted to address the streamflow forecasting problem, however, few methods are developed to make a systematic research. The objective of this study is to understand the underlying trend of streamflow so that a regression model can be developed to forecast the flow volume. In this paper, a hybrid streamflow forecast framework is proposed that integrates factor analysis, time series decomposition, data regression, and error suppression. Correlation coefficients between the current streamflow and the streamflow with lags are analyzed using autocorrelation function (ACF), partial autocorrelation function (PACF), and grey correlation analysis (GCA). Support vector regression (SVR) and generalized regression neural network (GRNN) models are integrated with seasonal and trend decomposition to make monthly streamflow forecast. Auto-regression and multi-model combination error correction methods are used to ensure the accuracy. In our experiments, the proposed method is compared with a stochastic autoregressive integrated moving average (ARIMA) streamflow forecast model. Fourteen models are developed, and the monthly streamflow data of Shigu and Xiangjiaba, China from 1961 to 2009 are used to evaluate our proposed method. Our results demonstrate that the integrated model of grey correlation analysis, Seasonal-Trend Decomposition Procedure Based on Loess (STL), Support Vector Regression (GCA-STL-SVR) exhibits an improved performance for monthly streamflow forecast. The average error of the proposed model is reduced to less than one-tenth in contrast to the state-of-the-art method and the standard deviation is also reduced by more than 30%, which implies a greater consistency.

## 1. Introduction

Accurate streamflow forecasting is of significant importance for planning and management of water resources, as well as early warning and mitigation of natural disasters such as droughts and floods (Yu et al., 2018). Nevertheless, affected by complex factors including precipitation, evaporation, runoff yield and confluence, topography and human activities, it is still challenging to achieve accurate streamflow forecasting (Senthil Kumar et al., 2013).

Until now, a large variety of streamflow forecast models have been proposed, mainly classified as physical and data-driven models. Physical models are good at providing insight into catchment processes, while they have been criticized for being difficult to implement. In contrast, data-driven models have minimum information requirements and rapid development times. Data-driven stochastic models have been used for streamflow forecasting. Autoregressive integrated moving average models (ARIMA) and its variants are widely used (Papacharalampous et al., 2018). While stochastic models are often limited by assumptions of normality, linearity and variable independence (Chen et al., 2018; Chen and Singh, 2018), the second data-driven type, machine learning shows a strong deep learning ability and extremely suitable for simulating the complex process. Over the past 50 years, research on machine learning has evolved from the efforts of a handful of computer engineers (Mitchell, 2006; Yuan and Abouelenien, 2015; Yuan et al., 2018). Artificial Neural Network (ANN) is a widely used method for long-term simulation and forecast (Aksoy and Dahamsheh, 2009; Moeeni and Bonakdari, 2016; Wu and Chau, 2010; Wu et al., 2009). But ANN still has some intrinsic disadvantages, such as slow convergence speed, less generalizing performance, arriving at a local minimum and over-fitting problems. Support vector machine (SVM) is based on the VC-dimension theory and structural risk minimization of statistical learning (Cortes and Vapnik, 1995). It transforms the problem into a quadratic optimization problem, theoretically, which can get the globally optimal solution, and solve the practical problems such as small sample, nonlinear, high dimension and local minimum (Smola and Lkopf, 2004; Vapnik, 2010). Maity et al. (2010) pointed out that SVR machine learning approaches were more popular due to their inherent advantages over traditional

https://doi.org/10.1016/j.jhydrol.2018.10.064
Received 29 April 2018; Received in revised form 22 October 2018; Accepted 27 October 2018
Available online xxx
0022-1694/ 

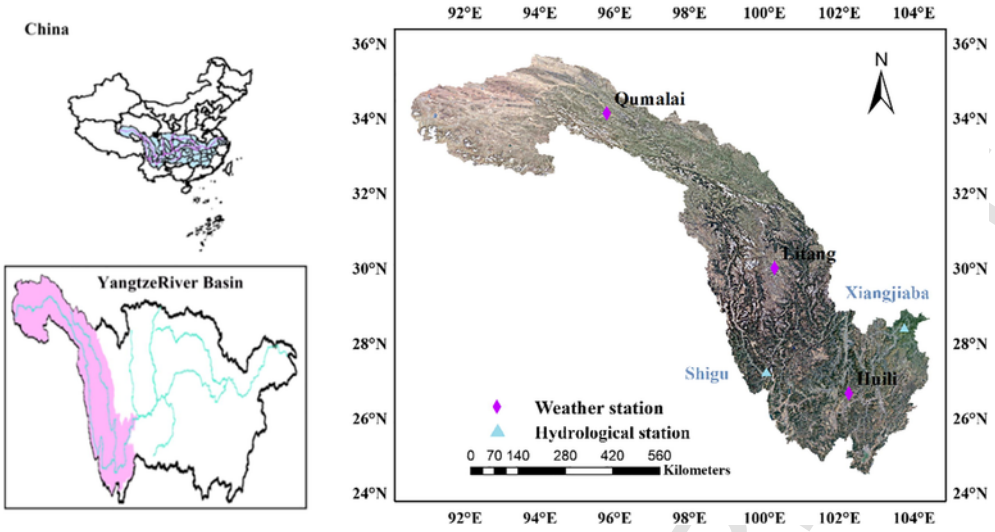


**Fig. 1.** The schematic of the Jinsha River and the gauging stations.

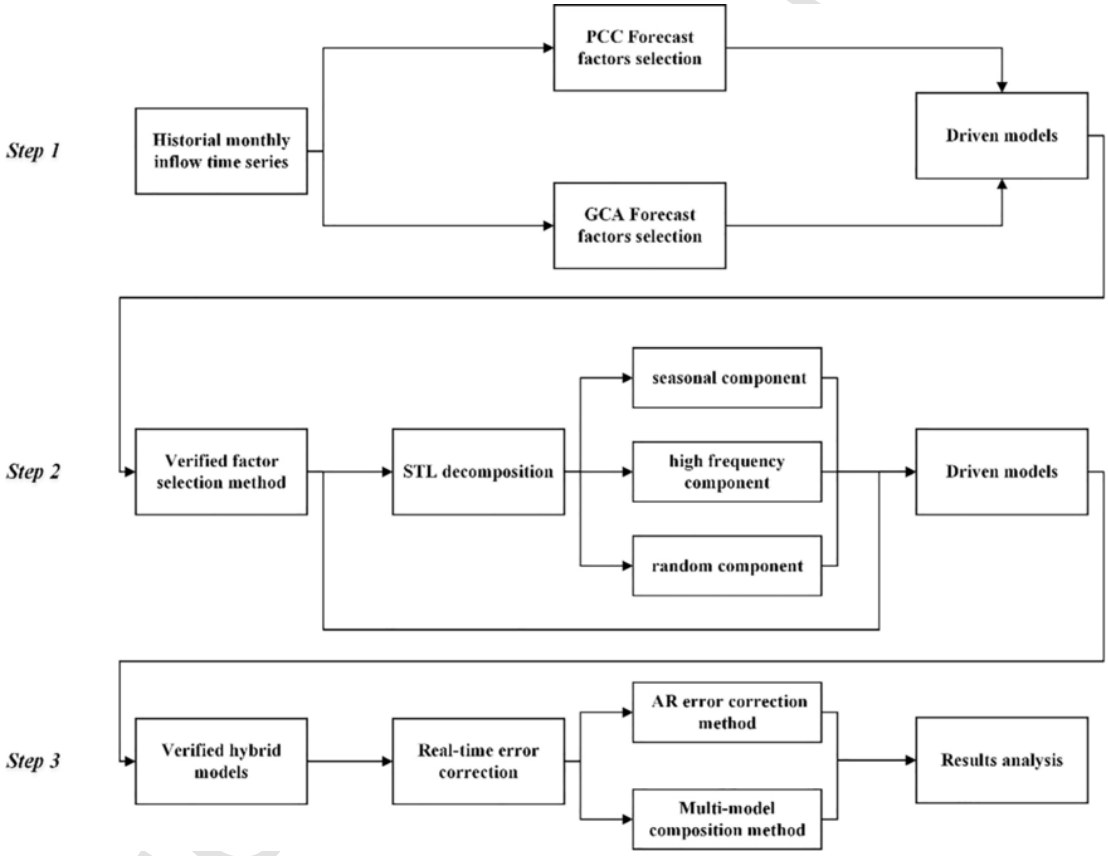


**Fig. 2.** Streamflow forecast framework.

modeling techniques. Kalteh (2015) employed genetic algorithm-support vector regression (GA-SVR) models for forecasting monthly flow on two rivers and obtained good performance. Papacharalampous et al. (2017) conducted large-scale computational experiments to compare stochastic and ML methods regarding their multi-step ahead forecasting properties and suggested that the ML methods exhibit a good performance. An important step of ANN and SVR models is to determine the significant input variables (Bowden et al., 2005a; Bowden et al., 2005b), some of which are correlated, noisy, and some input variables are less informative (Bowden et al., 2005a; Chen et al., 2013). Grey correlation analysis (GCA) evaluates the complex phenomena affected by many factors and a good metric to quantify the degree of association between the forecasting factors and the streamflow.

However, in any streamflow forecast model, there are three types of uncertainty caused by a number of factors: input uncertainty, model structure uncertainty and parameter uncertainty (Liu and Gupta, 2007). In order to reduce uncertainty and improve accuracy, a proven method, time series decomposition has been employed in lots of researches. Time series decomposition has the ability to analyze stream-

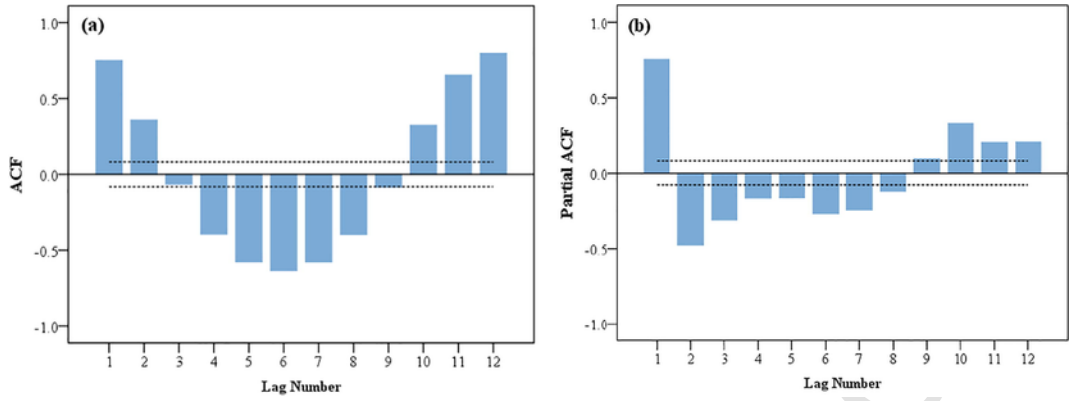


**Fig. 3.** Correlation analysis using ACF and PACF. (a) ACF with respect to the lag number. (b) PACF with respect to the lag number.

**Table 1**
Forecast error statistics of GCA-SVR, GCA-GRNN, PCC/ACF-SVR, and PACF-SVR.

| | MAPE | ≤5% | ≤10% | ≤20% | ≤30% | DC |
|---|---|---|---|---|---|---|
| GCA-SVR | 13% | 24% | 52% | 76% | 91% | 0.86 |
| GCA-GRNN | 17% | 14% | 38% | 71% | 84% | 0.82 |
| PCC/ACF-SVR | 16% | 23% | 45% | 72% | 88% | 0.83 |
| PACF-SVR | 16% | 22% | 41% | 71% | 87% | 0.82 |

flow temporal and spatial variation and extracting useful information as much as possible (Kisi, 2011). Wavelet analysis (Mallat, 1989) is developed on the basis of Fourier analysis. It is a local transformation of space and frequency. By stretching and translating, the signal can be multi-scale analyzed, therefore, it is suitable for the analysis of non-stationary hydrological time series (Guo et al., 2011; Kisi, 2011; Liu et al., 2014). Seasonal and trend decomposition using loess (STL) uses a locally weighted regression that enables processing of any type of seasonal variation data (Cleveland and Cleveland, 1990). Rojo et al. (2017) predicted airborne pollen series based on the seasonal and residual (stochastic) components of data series decomposed by using STL. Lafare et al. (2016) used STL to understand groundwater behavior in the Permo-Triassic Sandstone aquifer.

Different from time series decomposition, real-time error correction is a post-process method to reduce uncertainty and improving accuracy. The Kalman filter updating method can reflect various hydrological and hydraulic flow-fields by updating model input or parameters, but it tends to require a long computational time (Wu et al., 2012). Wu et al. (2012) tested the Kalman filter with simple neural networks and autoregressive models and pointed out that the results are similar. In addition, the multi-model combination approach advocates the synchronous use of the simulated discharges of a number of models to produce an overall integrated result which can be used as an alternative to that produced by a single model (Chen et al., 2015). Shamseldin et al. (1997) first introduced the multi-model combination concept into the hydrologic field. Since then there have been several more studies which have dealt with a multi-modal combination of hydrological models (Coulibaly et al., 2005; Wu et al., 2015; Xiong et al., 2001). Wu et al. (2015) proposed three coupling forecast methods which included real-time correction-combination forecast method, combination forecast real-time correction method for the purpose of improving the precision of flood forecasting.

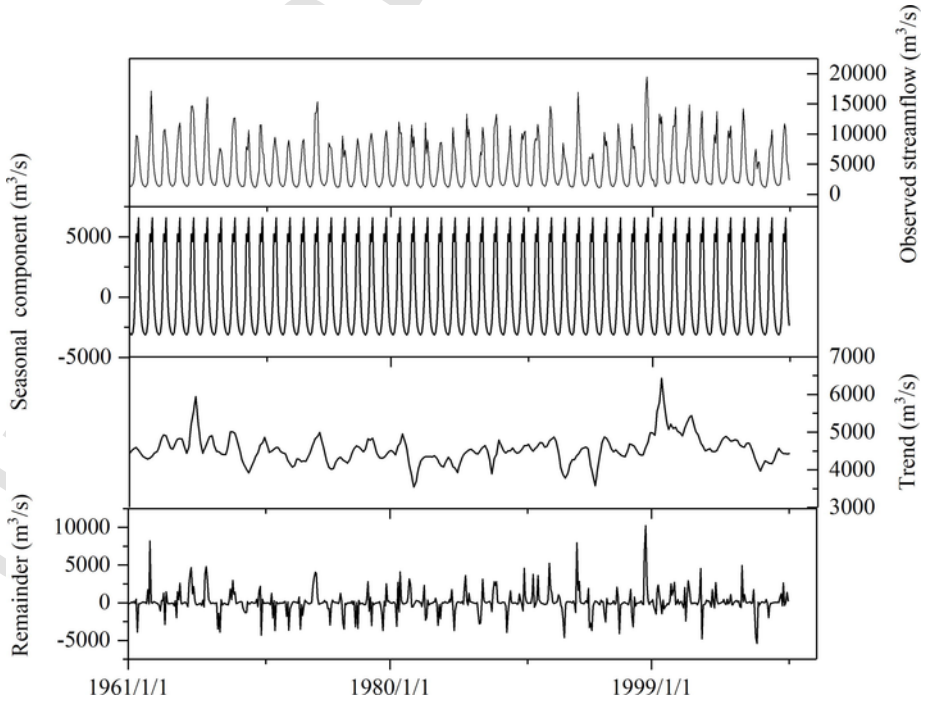


**Fig. 4.** Seasonal, trend and remainder components decomposed by STL.

**Table 2**
Hybrid streamflow forecast models, inputs, and output.

| Model | Input | Output |
|---|---|---|
| SVR/ANN | $Q_{t-1}$, $Q_{t-11}$, $Q_{t-12}$ | $Q_t$ |
| STL-SVR/STL-ANN | $Q(\mathrm{STL})_{t-1}$, $Q(\mathrm{STL})_{t-11}$, $Q(\mathrm{STL})_{t-12}$ | $Q_t$ |

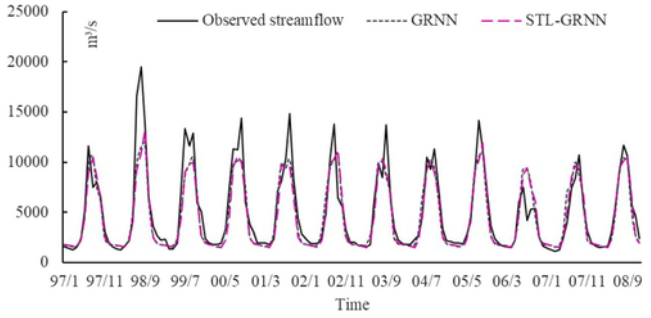


**Fig. 5.** Forecasted streamflow in the test period by using GRNN and STL-GRNN.

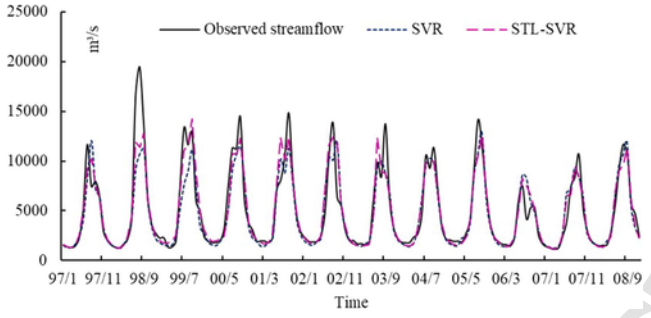


**Fig. 6.** Forecasted streamflow in the test period by using SVR and STL-SVR.

Many studies have been done to improve streamflow forecast accuracy. As described above, there exist comparative studies. For example, comparison of driving models (Kalteh, 2015; Moeeni and Bonakdari, 2016), predicting factor screening methods (Chen et al., 2013), multiple sequence decomposition (Guo et al., 2011; Kisi and Cimen, 2011; Liu et al., 2014) and multiple error correction methods (Chen et al., 2015; Wu et al., 2015).

A model requires a systematic integration of many components including factor analysis, time series decomposition, data regression, and error suppression, which enables accurate modeling of a hydro-system. In this paper, ANN and Support Vector Regression (SVR) are employed as regression models. SVR is more accurate but less efficient than the least squares support vector machine LSSVR (Suykens et al., 2002) because SVR solves a convex quadratic programming (CQP) problem to determine the regression and LSSVR just solves a set of reformulated linear equations. Generalized regression neural network is relatively simple in structure and training of network, there is no need to estimate the number of hidden layers and the number of hidden cells in advance, and there is an advantage of global convergence. The autocorrelation function (ACF), partial autocorrelation function (PACF) and grey correlation analysis (GCA) are candidate predictors screening methods. We develop hybrid models of SVR and GRNN for monthly streamflow forecast that couples with seasonal and trend decomposition. The error correction approaches are used to enhance the performance of the proposed hybrid models. The following article structure is organized as follows. Section 2 is the theory and methods. Section 3 is a description of the study area and data. Section 4 presents a case study. Section 5 concludes this paper with a summary.

## 2. Methods

The proposed streamflow forecasting framework consists of forecast factors selection, time series decomposition, model learning, and real-time error correction. Correlation coefficients between the current streamflow and the streamflow with N-month lag are analyzed using ACF, PACF, and GCA. The antecedent streamflow with a greater correlation is included in the input sets. Support vector regression (SVR) and generalized regression neural network models are integrated with seasonal and trend decomposition STL to forecast the monthly streamflow.

### 2.1. Support vector machine

Support vector machine (SVM), which is known as classification and then extended for regression, was proposed by (Vapnik, 1995). SVM is built based on the principle of the structural risk minimization rather than the empirical risk minimization. Support vector regression (SVR) is used to solve the problem of regression with SVM. The following is a brief description of SVR.

Suppose $N$ samples data for training are $\{(X_i, d_i)\}_i^N$, $X_i$ is the input vector, and $d_i$ means desired output. SVR results in

$$y = \boldsymbol{W}\varphi(X) + b \tag{1}$$

where $\varphi(X)$ is a non-linear mapping, $\mathbf{W}$ is a hyperplane, and $b$ is offset.

A penalty function is used in SVR:

$$\begin{cases} |d_i - y_i| \leqslant \varepsilon, & \textit{not allocating a penalty} \\ |d_i - y_i| > \varepsilon, & \textit{allocating a penalty} \end{cases} \tag{2}$$

When the estimated value is within the $\varepsilon$- insensitive tube, the loss value will be zero. Parameters of the regression function can be acquired by minimizing the following objective function:

$$\min\left[\frac{1}{2}\|W\|^2 + C\sum_{i=1}^{N} L^{\varepsilon}(y_i, d_i)\right] \tag{3}$$

$$L^{\varepsilon}(y_i, d_i) = \max(0, |y_i - d_i| - \varepsilon) \tag{4}$$

where $C$ represents the regularized constant that weighing the model complexity and the empirical error. A relative importance of the empirical risk will increase when the value of $C$ increases.

The slack variables $\xi^+$ and $\xi^-$ are introduced in Eq. (3) for the existence of fitting errors, the optimization problem of SVR will be as:

$$\min\left[\frac{1}{2}\|W\|^2 + C\sum_{i=1}^{N}(\xi_i^+ + \xi_i^-)\right] \tag{5}$$

Subject to: $(\boldsymbol{W}\varphi(\boldsymbol{X}_i) + b) - d_i \leqslant \varepsilon + \xi_i^+$

$d_i - (\boldsymbol{W}^T\varphi(\boldsymbol{X}_i) + b) \leqslant \varepsilon + \xi_i^-$

$\xi^+ > 0,\ \xi^- > 0$

Then use Lagrange multipliers to solve the above optimization problem in its dual form:

$$\max\left[\sum_{i=1}^{N}(\alpha_i^+ - \alpha_i^-)y_i - \varepsilon\sum_{i=1}^{N}(\alpha_i^+ + \alpha_i^-) - \frac{1}{2}\sum_{i=1,j=1}^{N}(\alpha_i^+ - \alpha_i^-)(\alpha_j^+ - \alpha_j^-)K(X_i, X_j)\right] \tag{6}$$

Subject to: $0 \leqslant \alpha_i^+ \leqslant C$

$0 \leqslant \alpha_i^- \leqslant C$

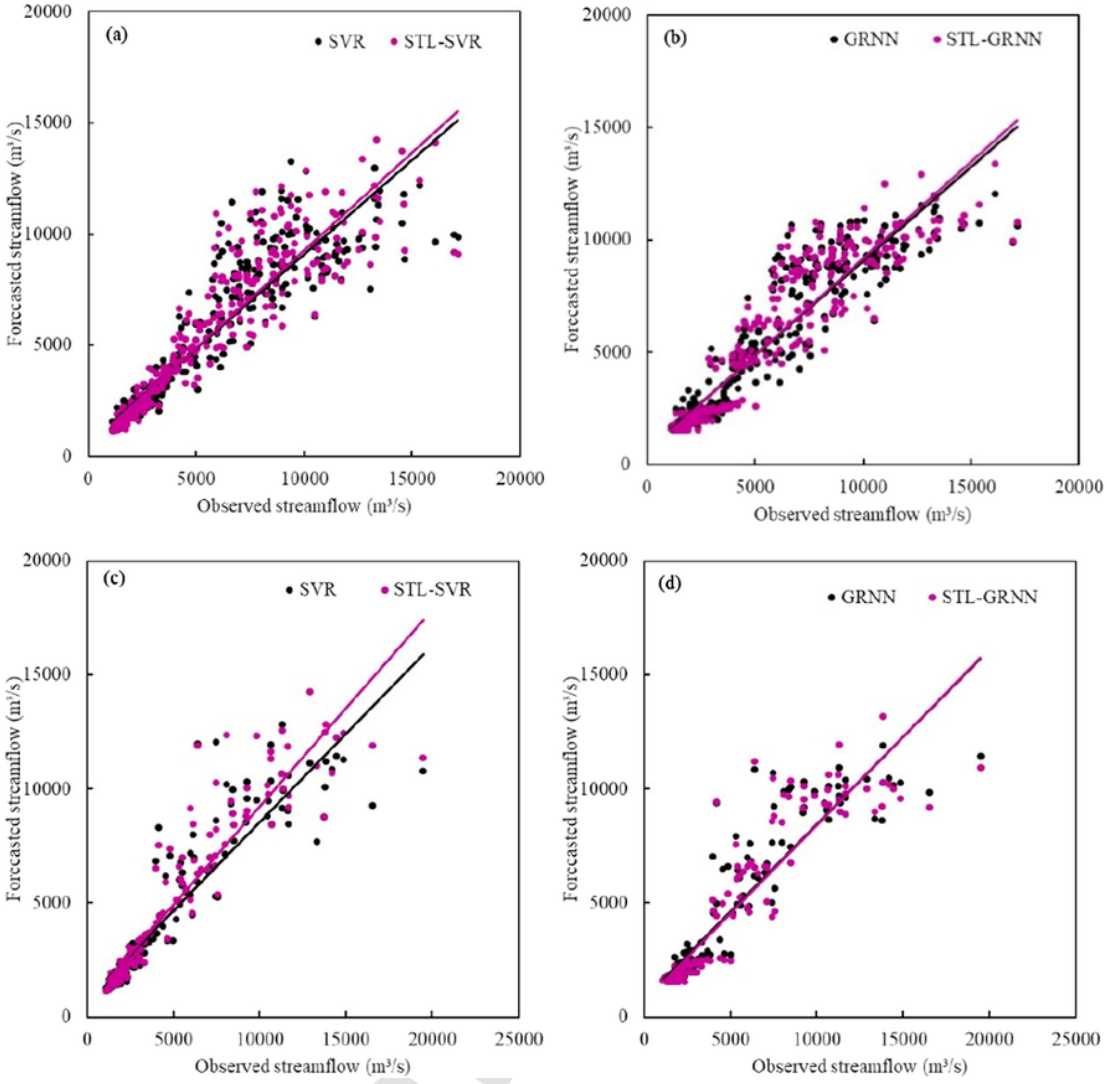


**Fig. 7.** Streamflow forecast error analysis in the training and test period. (a), (b) training period. (c), (d) test period.

**Table 3**
Forecast error statistics of models coupled with STL in both training and test period.

| Model | Training period | | Testing Period | |
|---|---|---|---|---|
| | DC | RMSE | DC | RMSE |
| SVR | 0.86 | 1354 | 0.82 | 1694 |
| STL-SVR | 0.87 | 1299 | 0.87 | 1433 |
| GRNN | 0.88 | 1277 | 0.83 | 1665 |
| STL-GRNN | 0.89 | 1221 | 0.82 | 1713 |

$$\sum_{i=1}^{N}(\alpha_i^+ - \alpha_i^-) = 0$$

$$K(X_i, X_j) = \varphi(\boldsymbol{X}_i)^T \cdot \varphi(\boldsymbol{X}_j) \tag{7}$$

$K(X_i, X_j)$ is a nonlinear kernel function, which can map the lower dimension input into a higher dimension linear space. Radial basis kernel function is used in this study.

### 2.2. Generalized regression neural network

Generalized regression neural network (Zaknich, 2013) is a special radial basis function neural network. Compared with the widely used BP neural network, GRNN has the following advantages: 1. Structure of GRNN neural network is relatively simple. In addition to the input and output layers, there are only two hidden layers, the pattern layer, and the summation layer, and the number of hidden neurons in the pattern layer is the same as the number of training samples. 2. Training of network is simple, the network training is completed once the training sample through the hidden layer. 3. There is no need to estimate the number of hidden layers and the number of hidden cells in advance. 4. Global convergence of GRNN.

The theoretical basis of the GRNN neural network is nonlinear regression analysis. Let the joint probability density function of the random variable $x$ and $y$ be $f(x, y)$ when the observed value of $x$ is $X$, the conditional expectation of $y$ for $X$ is:

$$\hat{y} = E(y|X) = \frac{\int_{-\infty}^{+\infty} yf(X, y)dy}{\int_{-\infty}^{+\infty} f(X, y)dy} \tag{8}$$

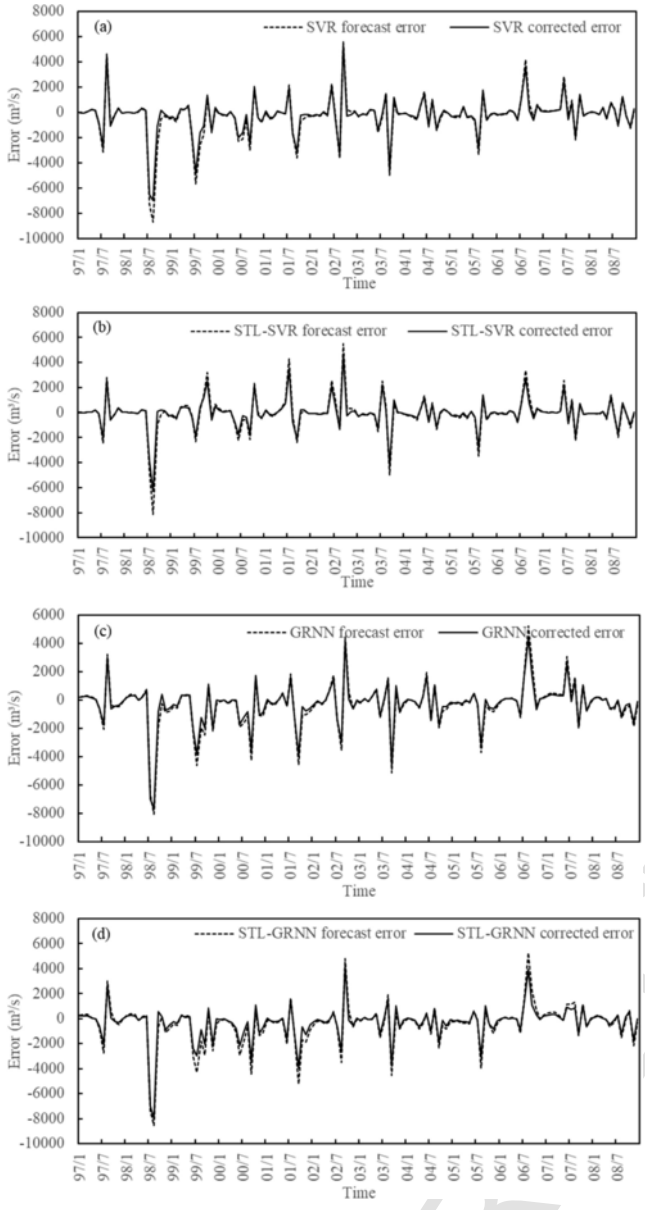


**Fig. 8.** AR corrected SVR, STL-SVR, GRNN and STL-GRNN forecast errors in the test period.

**Table 4**
Forecast error statistics before and after correction in the test period.

| | Before correction | | After correction | |
|---|---|---|---|---|
| | DC | RMSE | DC | RMSE |
| SVR | 0.82 | 1694 | 0.82 | 1490 |
| STL-SVR | 0.87 | 1433 | 0.90 | 1189 |
| GRNN | 0.83 | 1665 | 0.82 | 1618 |
| STL-GRNN | 0.82 | 1713 | 0.81 | 1660 |
| Combination | – | – | 0.77 | 1620 |

**Table 5**
Forecast errors of AR corrected GCA-STL-SVR and ARIMA model.

| | | AR corrected GCA-STL-SVR | ARIMA |
|---|---|---|---|
| **SD** | | 1196 | 1753 |
| **Mean** | | −47 | −510 |
| **Percentiles** | **25** | −272 | −740 |
| | **50** | −24 | −247 |
| | **75** | 196 | 151 |

Set the sample data set is $\setminus\{X_i, y_i\}$, $i = 1,2,\cdots,n$, the dimension of $X_i$ is $m$, the nonparametric estimates of probability density function $f(X, y)$ are as follows:

$$\hat{f}(X,y) = \frac{1}{n(2\pi)^{(m+1)/2}\sigma^{m+1}} \sum_{i=1}^{n} \exp\left[-\frac{(X-X_i)^T(X-X_i)}{2\sigma^2}\right] \exp\left[-\frac{(y-y_i)^2}{2\sigma^2}\right] \quad (9$$

Combining Eqs. (9) and (8) to get

$$\hat{y} = E(y|X) = \frac{\sum_{i=1}^{n} \exp\left[-\frac{(X-X_i)^T(X-X_i)}{2\sigma^2}\right] \int_{-\infty}^{+\infty} y \exp\left[-\frac{(y-y_i)^2}{2\sigma^2}\right] dy}{\sum_{i=1}^{n} \exp\left[-\frac{(X-X_i)^T(X-X_i)}{2\sigma^2}\right] \int_{-\infty}^{+\infty} \exp\left[-\frac{(y-y_i)^2}{2\sigma^2}\right] dy} \quad (10)$$

Hence, the estimation becomes

$$\hat{y} = \frac{\sum_{i=1}^{n} y_i \exp\left[-\frac{(X-X_i)^T(X-X_i)}{2\sigma^2}\right]}{\sum_{i=1}^{n} \exp\left[-\frac{(X-X_i)^T(X-X_i)}{2\sigma^2}\right]} \quad (11)$$

When the observed value of $x$ is $X$, the conditional expectation estimate of $y$ for $X$ is the weighted average of all sample observations $y_i$, the weight of $y_i$ is $\exp\left[-\frac{(X-X_i)^T(X-X_i)}{2\sigma^2}\right]$. The smoothing factor $\sigma$ needs to be optimized, for which we adopt cross-validation.

### 2.3. Grey correlation analysis

Grey system theory was proposed in the 1980s based on the mathematical theory of systems engineering (Ju-Long, 1982). Since then, the theory has become quite popular with its ability to deal with the systems that have partially unknown parameters. As a superiority to conventional statistical models, grey models require only a limited amount of data to estimate the behavior of unknown systems.

Grey correlation analysis is an important part of grey system theory. Grey correlation analysis method can evaluate the complex phenomena affected by many factors from the overall concept. In this paper, we introduce the grey correlation analysis to quantify the degree of association between the primary forecasting factor and the predicted runoff. The calculation process of grey correlation analysis is as follows:

Let $X_0 = \setminus\{x_0(1), \cdots, x_0(n)\}$ be the system characteristic behavior sequence, $X_i = \{x_i(1), \cdots, x_i(n)\}$ and $i = 1,2,\cdots,m$ is the relevant factor sequence. First of all to perform dimensionless processing on time sequence, and the initial values of each sequence are as follows:

$$X_i' = \frac{X_i}{x_i(1)} = (x_i'(1), x_i'(2), \cdots x_i'(n)) \quad (12)$$

where $i = 1,2,\cdots,m$.

Let $\Delta_i(k) = |x_0'(k) - x_i'(k)|$, the difference sequence of the initial value is:

$$\Delta_i = (\Delta_i(1), \Delta_i(2), \cdots, \Delta_i(n)) \quad (13)$$

Let $M = \max_i \max_k \Delta_i(k)$, $m = \min_i \min_k \Delta_i(k)$, the correlation coefficient between the target variable and related factors at each time is calculated as follows:

$$\gamma_{0i}(k) = \frac{m + \xi M}{\Delta_i(k) + \xi M} \tag{14}$$

where $\xi \in (0, 1)$, $k = 1,2,\cdots,n$ and $i = 1,2,\cdots,m$.

Average the correlation coefficients at each moment to systematically compare the degree of association between the target sequence and the correlation factor sequence. The correlation formula is as follows:

$$\gamma_{0i} = \frac{1}{n}\sum_{k=1}^{n}\gamma_{0i}(k) \tag{15}$$

### 2.4. STL decomposition

Seasonal and Trend decomposition using Loess (STL) uses the robust local weighted regression as a smoothing method to decompose the time series into seasonal, trend and residual items.

$$Y_v = Trend_v + Seasonal_v + Residual_v \tag{16}$$

where $Y_v$ is the time series observed value of the $v$ period, the trend component $Trend_v$ is considered low frequency, seasonal components. $Seasonal_v$ is considered to be high-frequency changes caused by seasonal interference, the remaining amount $Residual_v$ is a random component.

Loess is a local polynomial regression, a common method for smoothing two-dimensional scatterplots. The regression is done using the weighted least squares method, that is, the closer the data is to the estimated point, the greater the weight. Finally, the local regression model is used to estimate the value of the response variable. In this way, the whole fitting curve is obtained by the point-by-point operation. Loess is a nonparametric learning method defined as follows:

$$J(\theta) = \sum_{i=1}^{M} w_i[y_i - \theta^T f_i]^2 \tag{17}$$

where $f$ is the point to be estimated, $f_i$ is the sample point, $w_i$ is the weight $w^i = e^{\frac{(f^i - f)^2}{2k^2}}$.

The core of the STL algorithm is the iterative process of Loess, the iterative decomposition process is as follows:

(1) Set the initial value $k = 0$, $T_v^k = 0$.
(2) Remove trend items $Y_v - T_v^k$.
(3) The Loess smoothing is performed on the subsequence to obtain the time series $C_v^{k+1}$.
(4) Perform three times moving average on $C_v^{k+1}$ for lengths of $n_p$,$n_p$, 3, perform a Loess process to get the time series $L_v^{k+1}$, remove periodic differences.
(5) Remove trend items $S_v^{k+1} = C_v^{k+1} - L_v^{k+1}$.
(6) Remove season items $Y_v - S_v^{k+1}$.
(7) The Loess smoothing is performed on the subsequence $Y_v - S_v^{k+1}$ to obtain the time series $T_v^{k+1}$.
(8) Check whether $T_v^{k+1}$ convergence. If convergence, $S_v = S_v^{k+1}$, $T_v^{k+1}$, $R_v = Y_v - S_v - T_v$, if not, repeat the process (2)–(8).

### 2.5. Real-time error correction

#### 2.5.1. AR error correction method

The discrepancy between the model-predicted discharge and the actually observed past discharge is defined as an error which can be used as information for correction. If this error signal has a correlation, it can probably be used for improved prediction. With a time-series model of the error signal, an improved discharge forecast can be made by adding the error term to the previous model results. In this study, the error term was estimated using an autoregressive (AR) model which can be expressed as:

$$e_t = \sum_{k=1}^{p}\theta_i e_{t-k} + \xi \tag{18}$$

where $e$ is the streamflow forecasting error time series; $p$ represents the order of the autoregressive model; $\theta_i$ are the parameters of the autoregressive model, and $\xi$ is a pure white noise sequence having variance $\sigma^2$. Order selection criteria were used to determine the appropriate order.

#### 2.5.2. Multi-model composition method

Estimates of $N$ streamflow forecast models for the $t$-th period of time is $\hat{Q}_{it} i = 1,2,\cdots,N$, a combined estimate $Q_{ct}$ is defined

$$Q_{ct} = \sum_{i=1}^{N} w_i \hat{Q}_{it} + \xi_t \tag{19}$$

where $w_i$ is the weight assigned to the $i$-th model, $\sum_{i=1}^{N} w_i = 1$; and $\xi_t$ is the combination error term.

In order to obtain the weights $w_i$, an objective function is described as follows:

$$MinE(Q_{ct} - Q_{obs,t})^2 = MinE(\sum_{i=1}^{N} w_i \hat{Q}_{it} - -Q_{obs,t})^2 \tag{20}$$

and the Lagrange multiplier is used to solve the above problem.

## 3. Study area and data

Our study takes the Jinsha River as the study area. The Jinsha River is located in the upper reaches of the Yangtze River (China), with the basin area of 473,200 km$^2$, accounting for 26% of the Yangtze River Basin area. It has a total length of 3479 km, a natural drop of 5100 m. It is rich in hydropower resources and plays a vital role in economic development and ecological environmental conservation of China. Twenty-five hydropower dams in the Jinsha watershed are and will be constructed, which take the responsibility of flood control, agricultural hydroelectric power generation, and municipal and industrial water supply. With the completion of these hydropower dams, the Jinsha River hydropower resources are effectively developed and utilized. Discharge forecasting is significant to the optimal operation of these dams. This study focuses on the Shigu and Xiangjiaba gauging stations, which are hydrological control station of the upper reach and the lower reach of Jinsha River, respectively. A schematic of the Jinsha River and the gauging stations is given in Fig. 1.

We obtained the available quality-controlled and partially infilled daily streamflow (m$^3$/s) data at the Xiangjiaba (1961–2008) and Shigu (1970–2009) hydrologic stations, provided by the Yangtze River Waterway Bureau, China. Monthly streamflow data needed in this research were aggregated from daily data.

## 4. Case study

### 4.1. Frameworks of proposed models

The steps of our proposed streamflow forecast framework are listed as follows:

Step 1) ACF, PACF and GCA are used to select forecast factors.

The correlation coefficients between the current streamflow and the streamflow with N-month lag is calculated using ACF, PACF, and GCA.

Each antecedent monthly streamflow has three correlation coefficient values. The antecedent streamflow with the higher ACF correlation value is included to get an ACF input set. Similarly, we get PACF input set and GCA input set with the higher PACF and GCA value. Support vector machine (SVR) and generalized regression neural network are used as regression models. The statistical indicators evaluating the accuracy of prediction are mean average percentage error (MAPE), a proportion that errors less than 5%, 10%, 20%, 30% and deterministic coefficient (DC). Continuous monthly streamflow data from 1970 to 2009 are used in Shigu case study. Then the better input set can be determined by their forecast performances. Then the better input set can be determined by their forecast performances.

Step 2) On the basis of Step 1, we developed hybrid SVR and GRNN monthly streamflow forecast models coupling with seasonal and trend decomposition methods STL.

SVR and GRNN are used as the model, the better method proved in Step 1 is used to make forecast factors selection. The original sequences are decomposed into multiple sub-series by time series pre-processing techniques STL decomposition. Four models, SVR, STL-SVR and GRNN, STL-GRNN are built to forecast monthly streamflow of Xiangjiaba. The performances of the models are compared using the root-mean-square error (RMSE) and deterministic coefficient (DC).

Step 3) Two real-time error correction methods (the AR model and multi-model combination method) are used to enhance the performance.

A 3-order AR model is used for error correction. The parameters of the AR model are optimized according to the least-square method. As a comparison, the multi-model composition method is also used for the error correction. We train SVR and GRNN models. The framework is illustrated in Fig. 2.

### 4.2. Performance metrics

The statistical indicators evaluating the accuracy of prediction are mean average percentage error (MARE), the proportion that errors less than 5%, 10%, 20%, 30% and deterministic coefficient (DC), and the root-mean-square error (RMSE). MAPE is calculated according to Eq. (21). An accurate model has the MAPE metric value close to 0. RMSE is a frequently used measure of the differences between values predicted and the values actually observed. RMSE represents the sample standard deviation of the differences between predicted values and observed values. RMSE is calculated according to Eq. (22). DC is the proportion of the variance in the dependent variable that is predictable from the independent variables. It provides a measure of the quality of outcomes replicated by the model, based on the proportion of total variation of outcomes explained by the model DC is calculated according to Eq. (23)

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|Y_i^{obs} - Y_i^{est}\right|}{Y_i^{obs}} \times 100\% \tag{21}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i^{obs} - Y_i^{est})^2} \tag{22}$$

$$DC = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{23}$$

### 4.3. Forecast factor selection

Forecasting factor and regression models are essential for developing a streamflow forecast framework. In this part, the SVR model and GRNN model, two machine learning method are built for data simulation. The autocorrelation function, the partial autocorrelation function, and grey correlation analysis are introduced to quantify the correlation degree between the streamflow and potential forecasting factors.

It is an effective forecast method by using the variable's own historical records to make an estimation. If $Q(t)$ is the streamflow to be forecasted, select antecedent streamflow $Q(t-1), Q(t-2), \cdots Q(t-i), \cdots, Q(t-12)$ as alternative factors, where $Q(t-i)$ means streamflow $i$ month ahead of forecast month. The grey correlation degree between the alternative factors $Q(t-i)$ and $Q(t)$ is calculated, results are as follows:

$$\gamma_{t-1} = 0.88, \gamma_{t-2} = 0.70, \gamma_{t-3} = 0.64, \gamma_{t-4} = 0.66, \gamma_{t-5} = 0.74, \gamma_{t-6} = 0.69$$
$$\gamma_{t-7} = 0.55, \gamma_{t-8} = 0.68, \gamma_{t-9} = 0.74, \gamma_{t-10} = 0.81, \gamma_{t-11} = 0.87, \gamma_{t-12} = 0.83$$

Therefore $Q(t-1), Q(t-5), Q(t-9), Q(t-10), Q(t-11), Q(t-12)$ are adopted as forecast factors, and SVR and GRNN models are used to describe the following function $Q(t) = f(Q(t-1), Q(t-5), Q(t-9), Q(t-10), Q(t-11), Q(t-12))$. The high correlation between the flow with 11 and 12-month lag and the current flow implies the annual streamflow variation; whereas the flow with 1-month lag usually has a relatively similar value to the current flow. Flows with a shorter lag such as 5, 6, and 7 months imply seasonal fluctuations as well as long-term changes from atmospheric circulation.

Pearson correlation coefficient (PCC), autocorrelation function, and partial autocorrelation function are used to analyze the streamflow time series. The autocorrelation and partial autocorrelation patterns of the Shigu streamflow are presented in Fig. 3. Fig. 3(a) shows ACF with respect to different lag numbers. It is clear that the streamflow fluctuates and our results demonstrate significant autocorrelations at the time lags 1, 5, 6, 7, 11, and 12 months. Fig. 3(b) shows the PACF with respect to the lag number. The time series exhibits significant partial autocorrelations at times lags 1, 2, 3, and 10 months. The results from PCC analysis suggest that streamflow has a high correlation with streamflow at time lags 1, 5, 6, 7, 11, and 12 months, which is highly similar to the results of ACF.

Four models GCA-SVR, GCA-GRNN, PCC/ACF-SVR, and PACF-SVR are devised to determine the forecast factors. The inputs of GCA-SVR and GCA-GRNN are streamflow with a lag of 1, 5, 9, 10, 11, 12 months. The inputs of the PCC/ACF-SVR model are streamflow with a lag of 1, 5, 6, 7, 11, 12 months. The inputs of PACF-SVR model are streamflow with a lag of 1, 2, 3, and 10 months.

Continuous monthly streamflow data from 1970 to 2009 are used in Shigu case study. In supervised learning, data sets are often divided into three sets, namely the training set, the testing set, and the validation set. When the sample size is small, it is common to divide data into training and testing set, and the cross-validation method is used. In our study, data from1970 to 1999 are used for training the model, which accounts for 75% of the total dataset. A 4-fold cross-validation method is applied. The remaining 25% of data (i.e., data from 2000 to 2009) are used for model validation.

The mean average percentage error (MAPE) of forecasting, the proportion of errors that are less than 5%, 10%, 20%, and 30%, and the deterministic coefficient (DC) are shown in Table 1. An accurate model has the MAPE metric value close to 0, DC value close to 1, and the dynamic range of error is small. The MAPEs of GCA-SVR and GCA-GRNN are 13% and 17%, respectively, which indicates that GCA-SVR performs better. For proportion that error less than 5%, 10%, 20% and 30%, GCA-SVR is 24%, 52%, 76% and 91%, GCA-GRNN is 14%, 38%, 71% and 84%. A large proportion of smaller error range implies that the errors are small. The DC of GCA-SVR and GCA-GRNN are 0.86 and 0.82, respectively. It demonstrates that learning model SVR is superior to GRNN in forecasting the streamflow of Jinsha River. When comparing GCA-SVR with PCC/ACF-SVR and PACF-SVR, SVR model cou-

pled with different input factors, it is found that the forecast performance of GCA-SVR is better than that of PCC/ACF-SVR and PACF-SVR.

### 4.4. Hybrid models based on STL decomposition

As time series pre-processing techniques are effective to improve the performance, STL is used to decompose the original streamflow data into multiple sub-series due to its advantages of being allowed to change over time and having the better robustness to the anomaly. Fig. 4 shows that the observed streamflow time series contains a stationary seasonal component with a period of 12 months, trend component increased significantly in the 1990s and decreased in 2000s under the combined effects of climate change and human activities. Four models, GCA-SVR, GCA-STL-SVR and GCA-GRNN, GCA-STL-GRNN are proposed in this paper to forecast monthly streamflow of Xiangjiaba. For Xiangjiaba hydrologic station, monthly streamflow data from 1961 to 2008, the first 36 years are used to model calibration, the rest 12 years are used to model evaluation.

Hybrid models and corresponding input and output are shown in Table 2. $Q_{t-1}$, $Q_{t-11}$, $Q_{t-12}$ represent original streamflow at 1, 11 and 12 months ahead forecasting month, $Q(\mathrm{STL})_{t-1}$, $Q(\mathrm{STL})_{t-11}$, and $Q(\mathrm{STL})_{t-12}$ represent sub-sequence of streamflow decomposed by STL. Model inputs are selected using GCA.

Forecast results of GRNN, STL-GRNN, SVR, and STL-SVR in test period are shown in Figs. 5 and 6. It can be seen that all models fit the observed streamflow well. Fig. 7 provides the forecast ability comparison between a single model and the hybrid model, it can be observed that hybrid models coupled with STL decomposition are better than models without decomposition. However, it is also obvious that models have better forecasts at conventional runoff samples, for flood caused by heavy rain, the above models are not so satisfying as more uncertainty exists in these extreme value. Table 3 gives error statistic results in both training period and test period. An analysis shows that SVR has a better streamflow forecast results than GRNN. STL decomposition methods improve the accuracy of prediction compared with the results of SVR and GRNN model. STL-SVR is the best model for monthly streamflow forecast on Jinsha River.

### 4.5. Real-time error correction analysis

We use the 3rd order AR model and its parameters are determined by minimizing the least square error function. The order of the AR model is determined using the Bayesian information criterion (BIC). For STL-SVR forecast errors, BICs are 15.99, 14.08, 14.01, 15.04 and 16.06 for orders of 1, 2, 3, 4, and 5, respectively. Hence, the 3rd order is used in our experiments.

Errors of AR corrected SVR, STL-SVR, GRNN, and STL-GRNN model are depicted in Fig. 8.

As a comparison, the multi-model composition method is also used for the error correction. Weight parameters of SVR and GRNN model are 0.78 and 0.24, respectively. Results of the AR error correction method and multi-model composition method in the test period are given in Table 4. DC and RMSE value indicate that the AR corrected models are better than the original SVR and STL-SVR model. For the GRNN model and STL-GRNN model, consistency measured with DC is a little decreased but whole errors measured with RMSE are improved obviously. AR corrected GCA-STL-SVR exhibits the best performance, multi-model composition method has no significant contribution in our study.

### 4.6. Comparison between AR-corrected GCA-STL-SVR model and stochastic ARIMA forecast model

AR-corrected GCA-STL-SVR model is compared with a stochastic autoregressive integrated moving average (ARIMA) streamflow forecast model. Table 5 gives the average error, standard deviation (SD), percentiles of 25 ($Q_1$), 50 ($Q_2$) and 75 ($Q_3$). The average error of AR corrected GCA-STL-SVR model is -47 with an SD of 1196; whereas the average error of Autoregressive integrated moving average (ARIMA) is -510 with an SD of 1753. It is clear that the error of our proposed method is much smaller and the spread of the error range is also small. For the ARIMA model, the negative errors account for a large amount, which means that the ARIMA forecast tends to underestimate the true runoff. It demonstrates that the AR corrected GCA-STL-SVR model exhibits a better performance for monthly streamflow forecast of Jinsha River.

## 5. Conclusions

The monthly runoff forecast in the Jinsha River Basin has received much attention in recent years. Due to the impact of the subtropical monsoon, inner-annual alterations of runoff are extremely complex and hard to be accurately forecasted. Most of the existing researches used machine learning model combined with time series decomposition to improve forecasting accuracy. In this paper, a more systematic forecasting method was researched. We analyzed ACF, PACF, and GCA correlation coefficients to determine the forecast factors STL decomposition used to divide the original sequence into trends, periodic items to understand the underlying trend of streamflow. Error real-time correction post-processing techniques were introduced to form a complete forecasting framework. Fourteen models were finally developed in this paper and AR corrected GCA-STL-SVR was regarded as the best model for streamflow forecast of Jinsha River.

First, four models GCA-SVR, ACF-SVR, PACF-SVR and GCA-GRNN were proposed. Experiments with continuous monthly streamflow data in Shigu shows that GCA-SVR is superior to ACF-SVR, PACF-SVR, and GCA-GRNN. The result proves the advantage of GCA, as widely used forecast factors selection approaches ACF and PACF measure the linear correlation of variables, with the drawback of ignoring the non-linear relation.

STL was used to decompose the original streamflow data into multiple sub-series due to its advantages of being allowed to change over time and having the better robustness to the anomaly. The observed streamflow time series contains a stationary seasonal component with a period of 12 months, an increased trend in the 1990s and a decreased trend in 2000s under the combined effect of climate change and human activities. Then another four models, GCA-SVR, GCA-STL-SVR and GCA-GRNN, GCA-STL-GRNN were proposed to forecast monthly streamflow in the lower reach of the Jinsha River. Forecast results show that hybrid models coupled with STL decomposition provide better forecasts than models without decomposition. GCA-STL-SVR is more effective for the purpose of improving the precision of monthly streamflow forecast.

Then the AR model and multi-model composition were applied to correct the forecast error of GCA-SVR, GCA-STL-SVR, GCA-GRNN, and GCA-STL-GRNN models. DC and RMSE value indicate that the AR corrected models are better than the original models without correction. For the GRNN model and STL-GRNN model, consistency measured with DC is a little decreased but whole errors measured with RMSE are improved obviously. Multi-model composition method has no significant contribution in our study. AR corrected GCA-STL-SVR model was developed as the streamflow forecast framework of Jinsha River.

A stochastic autoregressive integrated moving average (ARIMA) streamflow forecast model was implemented to evaluate the performance of the proposed AR corrected GCA-STL-SVR model. It is clear that the error of our proposed method is much smaller and the spread of the error range is also small. ARIMA forecast tends to underestimate the true runoff. It demonstrates that the AR corrected GCA-STL-SVR model exhibits a better performance for monthly streamflow forecast of Jinsha River.

## Acknowledgments

This work is supported by the National Natural Science Foundation of China (No: 51809242), and special thanks are given to the China scholarship council's support and the anonymous reviewers and editors for their constructive comments.

## References

Aksoy, H., Dahamsheh, A., 2009. Artificial neural network models for forecasting monthly precipitation in Jordan. Stochastic Environ. Res. Risk Assess. 23 (7), 917–931.

Chen, L., Singh, V., Huang, K., 2018. Bayesian technique for the selection of probability distributions for frequency analyses of hydrometeorological extremes. Entropy 20 (2), 117.

Chen, L., Singh, V.P., 2018. Entropy-based derivation of generalized distributions for hydrometeorological frequency analysis. J. Hydrol. 557, 699–712.

Chen, L., Ye, L., Singh, V., Zhou, J., Guo, S., 2013. Determination of input for artificial neural networks for flood forecasting using the copula entropy method. J. Hydrol. Eng. 19 (11), 217–226.

Chen, L., Zhang, Y., Zhou, J., Singh, V.P., Guo, S., Zhang, J., 2015. Real-time error correction method combined with combination flood forecasting technique for improving the accuracy of flood forecasting. J. Hydrol. 521, 157–169.

Cleveland, R.B., Cleveland, W.S., 1990. STL: A seasonal-trend decomposition procedure based on loess. J. Off. Stat. 6 (1), 3–33.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.

Coulibaly, P., Haché, M., Fortin, V., Bobée, B., 2005. Improving daily reservoir inflow forecasts with model combination. J. Hydrol. Eng. 10 (2), 91–99.

Guo, J., Zhou, J., Qin, H., Zou, Q., Li, Q., 2011. Monthly streamflow forecasting based on improved support vector machine model. Expert Syst. Appl. 38 (10), 13073–13081.

Ju-Long, D., 1982. Control problems of grey systems. Syst. Control Lett. 1 (5), 288–294.

Kalteh, A.M., 2015. Wavelet genetic algorithm-support vector regression (wavelet GA-SVR) for monthly flow forecasting. Water Resour. Manage. 29 (4), 1283–1293.

Kisi, O., 2011. Wavelet regression model as an alternative to neural networks for river stage forecasting. Water Resour. Manage. 25 (2), 579–600.

Kisi, O., Cimen, M., 2011. A wavelet-support vector machine conjunction model for monthly streamflow forecasting. J. Hydrol. 399 (1–2), 132–140.

Lafare, A.E.A., Peach, D.W., Hughes, A.G., 2016. Use of seasonal trend decomposition to understand groundwater behaviour in the Permo-Triassic Sandstone aquifer, Eden Valley, UK. Hydrogeol. J. 24 (1), 141–158.

Liu, Y., Gupta, H.V., 2007. Uncertainty in hydrologic modeling: toward an integrated data assimilation framework. Water Resour. Res. 43 (7), W07401.

Liu, Z., Zhou, P., Chen, G., Guo, L., 2014. Evaluating a coupled discrete wavelet transform and support vector regression for daily and monthly streamflow forecasting. J. Hydrol. 519, 2822–2831.

Maity, R., Bhagwat, P.P., Bhatnagar, A., 2010. Potential of support vector regression for prediction of monthly streamflow using endogenous property. Hydrol. Process. 24 (7), 917–923.

Mallat, S.G., 1989. A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. IEEE Trans. Pattern Anal. Mach. Intell. 11 (7), 674–693.

Mitchell, T.M., 2006. The Discipline of Machine Learning Vol. 9, Carnegie Mellon University, School of Computer Science, Machine Learning Department, Pittsburgh, PA.

Moeeni, H., Bonakdari, H., 2016. Forecasting monthly inflow with extreme seasonal variation using the hybrid SARIMA-ANN model. Stoch. Env. Res. Risk Assess. 31 (8), 1997–2010.

Papacharalampous, G., Tyralis, H., Koutsoyiannis, D., 2018. Predictability of monthly temperature and precipitation using automatic time series forecasting methods. Acta Geophys. 1–25.

Papacharalampous, G.A., Tyralis, H., Koutsoyiannis, D., 2017. Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes. J. Hydrol. 10.

Rojo, J., Rivero, R., Romeromorte, J., Fernándezgonzález, F., Pérezbadia, R., 2017. Modeling pollen time series using seasonal-trend decomposition procedure based on LOESS smoothing. Int. J. Biometeorol. 61 (2), 1–14.

Senthil Kumar, A.R., Goyal, M.K., Ojha, C.S., Singh, R.D., Swamee, P.K., 2013. Application of artificial neural network, fuzzy logic and decision tree algorithms for modelling of streamflow at Kasol in India. Water Sci. Technol. A J. Int. Assoc. Water Pollut. Res. 68 (12), 2521–2526.

Shamseldin, A.Y., O'Connor, K.M., Liang, G.C., 1997. Methods for combining the outputs of different rainfall–runoff models. J. Hydrol. 197 (1–4), 203–229.

Smola, A.J., Lkopf, B., 2004. A tutorial on support vector regression. Stat. Comput. 14 (3), 199–222.

Suykens, J.A.K., Gestel, T.V., Brabanter, J.D., Moor, B.D., Vandewalle, J., 2002. Least squares support vector machines. Int. J. Circuit Theory Appl. 27 (6), 605–615.

Vapnik, V., 2010. Statistical learning theory. DBLP 99–150.

Vapnik, V.N., 1995. The nature of statistical learning theory. Springer 988–999.

Wu, C.L., Chau, K.W., 2010. Data-driven models for monthly streamflow time series prediction. Eng. Appl. Artif. Intell. 23 (8), 1350–1367.

Wu, C.L., Chau, K.W., Li, Y.S., 2009. Predicting monthly streamflow using data-driven models coupled with data-preprocessing techniques. Water Resour. Res. 45 (8), 2263–2289.

Wu, J., Zhou, J., Chen, L., Ye, L., 2015. Coupling Forecast Methods of Multiple Rainfall-Runoff Models for Improving the Precision of Hydrological Forecasting. Water Resour. Manage. 29 (14), 5091–5108.

Wu, S.J., Lien, H.C., Chang, C.H., Shen, J.C., 2012. Real-time correction of water stage forecast during rainstorm events using combination of forecast errors. Stoch. Env. Res. Risk Assess. 26 (4), 519–531.

Xiong, L., Shamseldin, A.Y., O'Connor, K.M., 2001. A non-linear combination of the forecasts of rainfall-runoff models by the first-order Takagi-Sugeno fuzzy system. J. Hydrol. 245 (1–4), 196–217.

Yu, X., Zhang, X., Qin, H., 2018. A data-driven model based on Fourier transform and support vector regression for monthly reservoir inflow forecasting. J. Hydro-environ. Res. 18, 12–24.

Yuan, X., Abouelenien, M., 2015. A multi-class boosting method for learning from imbalanced data. Int. J. Granular Comput., Rough Sets Intell. Syst. 4 (1), 13–29.

Yuan, X., Xie, L., Abouelenien, M., 2018. A regularized ensemble framework of deep learning for cancer detection from multi-class, imbalanced training data. Pattern Recogn. 77, 160–172.

Zaknich, A., 2013. General regression neural network. Revue De Physique Appliquée iv (6), 1321–1325.