Seasonal streamflow forecasts using mixture-kernel GPR and advanced methods of input variable selection

Shuang Zhu, Xiangang Luo, Zhanya Xu and Lei Ye

ABSTRACT

Gaussian Process Regression (GPR) is a new machine-learning method based on Bayesian theory and statistical learning theory. It provides a flexible framework for probabilistic regression and uncertainty estimation. The main effort in GPR modelling is determining the structure of the kernel function. As streamflow is composed of trend, period and random components. In this study, we constructed a mixture-kernel composed of squared exponential kernel, periodic kernel and a rational quadratic term to reflect different properties of streamflow time series to make streamflow forecasts. A relevant feature-selection wrapper algorithm was used, with a top-down search for relevant features by Random Forest, to offer a systematic factors analysis that can potentially affect basin streamflow predictability. Streamflow prediction is evaluated by putting emphasis on the degree of coincidence, the deviation on low flows, high flows and the error level. The objective of this study is to construct a seasonal streamflow forecasts model using mixture-kernel GPR and the advanced input variable selection method. Results show that the mixture-kernel GPR has good forecasting quality, and top importance predictors are streamflow at 12, 6, 5, 1, 11, 7, 8, 4 months ahead, Nino $1+2$ at 11, 5, 12, 10 months ahead.

Key words | GPR, model uncertainty, relevant feature selection wrapper algorithm, streamflow prediction

Shuang Zhu
Xiangang Luo (corresponding author)
Zhanya Xu
Faculty of Information Engineering,
China University of Geosciences (Wuhan),
Wuhan 430074,
China
E-mail: *billlxg@126.com*

Lei Ye
School of Hydraulic Engineer Dalian University of Technology,
Dalian University of Technology,
Dalian,
China

INTRODUCTION

Accurate estimation of streamflow is vital for establishing agricultural water demands, hydroelectricity production needs, and environmental protection (Li *et al.* 2015; Xie *et al.* 2015), it is extremely important to investigate the accuracy of streamflow forecast models and confirm the most reliable one for a specific prediction condition. Streamflow forecasts can be performed using either physical/conceptual or data driven models. While physical/conceptual models are good at providing insight into catchment processes, they have been criticised for being difficult to implement in forecasting applications, requiring many different types of data and resulting in models that are overly complex (Beven 2006). In contrast, data driven models have minimum information requirements, rapid development times, and have been found to be accurate in various hydrological forecasting applications (Papacharalampous *et al.* 2018). Data driven stochastic models have traditionally been used for streamflow forecasting. Autoregressive integrated moving average (ARIMA) models have been the most widely used stochastic models for streamflow forecasting.

Stochastic models are linear models and are limited in their ability to forecast non-linear data. To effectively forecast non-linear data, researchers in the last two decades have increasingly begun to utilize machine learning (ML) to forecast hydrological data (El-Shafie *et al.* 2013; Yousefi *et al.* 2015; Karimi *et al.* 2017; Lu *et al.* 2017), as it provides a good solution for small sample, high dimension, and non-linear forecasting problems. Adamowski *et al.* (2012)

doi: 10.2166/nh.2018.023

investigated artificial neural network (ANN) water-demand prediction models using historical data of water demand, temperature and rainfall. Valipour *et al.* (2012) investigated the searching procedure to find the near-optimal structure of the developed ANNs that was used to predict monthly streamflow. Papacharalampous *et al.* (2017) conducted large-scale computational experiments to compare 11 stochastic to nine ML methods regarding their multi-step ahead forecasting properties and found the ML methods do not differ dramatically from the stochastic methods by quantifying the forecasting performance using 18 metrics. Therefore, wavelet decomposition techniques and hybrid models were studied to further improve prediction accuracy. Wu *et al.* (2009) examined using different hybrid models of moving average (MA), wavelet multi-resolution analysis (WMRA), and singular spectrum analysis (SSA) coupled with an ML model to improve the streamflow prediction accuracy. Zhu *et al.* (2016) conducted Yangtze River streamflow estimation using a support vector machine and several time series decomposition method.

However, for single point forecast model, simulation is of limited value because it merely indicates a single possible future value for the variable and does not convey information about the level of uncertainty, which is intrinsically associated with forecasting. Particularly for hydrological processes, after a period of growing hopes for a dramatic reduction in uncertainty in modelling, the value of stochastic modelling in dealing with uncertainty and risk is again appreciated (Koutsoyiannis *et al.* 2008; Montanari *et al.* 2013; Tyralis & Koutsoyiannis 2014). Gaussian Process Regression (GPR), which was originally formulated by Rasmussen and his coworkers, provides a 'principled, practical, and probabilistic approach to learning in kernel machines' (Rasmussen & Williams 2006). The advantage of GPR over many other machine learning methods lies in its seamless integration of hyperparameter estimation, model training, and uncertainty estimation; the results are less affected by subjectivity and more interpretable. Gaussian processes (GP) assume that the joint probability distribution of model outputs is Gaussian. In the hydrological literature, the notion of GP is underlying the Kriging algorithm in classical geostatistics, the autoregressive moving average (ARMA) models, Kalman filters, geostatistical inversion methods, and radial basis function (RBF) networks (Sun *et al.* 2014). GPR has a rigorous statistical learning theory foundation, the prediction is probabilistic so that one can compute empirical confidence intervals and decide based on those if one should refit the prediction in some region of interest. In addition, treat stochastic streamflow process as a generalization of a probability Gaussian process distribution and it turns out that the computations required for inference and learning are relatively easy. However, hydrologic records are generally characterized by a long-term persistence and the skewed distribution (Markonis & Koutsoyiannis 2015). Attempts have been made to adapt standard models to enable treatment of skewness. Koutsoyiannis (2000) proposed a generalized framework to incorporate short-memory (autoregressive moving average) and long-memory (fractional Gaussian noise) models, simultaneously, it explicitly preserves the coefficients of skewness of the processes.

The application of GPR is gaining popularity in many areas, but it is still rare in streamflow forecasting. Sun *et al.* (2014) applied GPR to probabilistic streamflow forecasting. Sun pointed out that the main effort in GPR modeling was determining the structure and hyperparameters of the kernel function. However, a commonly squared exponential kernel function was the only used kernel in their research. In this paper, we will build a mixture-kernel GPR streamflow forecast model and study its performance.

As streamflow is composed of trend, period and random components, the kernel used is composed of squared exponential kernel, periodic kernel and a rational quadratic term. Each term reflects different properties of the dataset. Squared exponential kernel explains a long term, smooth rising trend. A periodic covariance function with a period of one year to seasonal component models the seasonal variation. A rational quadratic term is used to model the small irregularities. The hyperparameter value of the kernel function has a great influence on learning and prediction. In this study, hyperparameter is optimized using gradient ascent on the log-marginal-likelihood. Four-fold cross-validation techniques are used to prevent overfitting. Following a normal distribution is the basic assumption of GPR, while streamflow distribution are highly skewed and their forecast errors are typically non-normal and display non-constant variance (Chen & Singh 2017). For non-Gaussian distributions of streamflow two strategies are commonly used. The first is to convert the process, using a nonlinear transformation, to a Gaussian process. The second strategy is to explicitly preserve, at the time scales of interest, higher order marginal moments, in particular the third moment or, equivalently, the skewness coefficient (Koutsoyiannis 2016). We adopt the Box-Cox transformation (Box & Cox 1964; George & Foster 2000) to normalize prediction variable errors and stabilize their variance in the transformed space.

The second purpose of this work is to offer a systematic analysis of factors that can potentially affect basin streamflow predictability. The study area selected is dominated by East Asian monsoon climate, runoff regime is extremely uneven and is affected by kinds of climate characteristics. The anomalous distribution of sea temperature has the characteristics of wide range, large thickness and long duration. It is often a foreboding to atmospheric circulation anomaly and can provide information for long-term hydrological forecasting. Guo *et al.* (2002) studied the relationships between floods in the Yangtze River valley and sea surface temperature anomalies (SSTA) in the Pacific and Indian Oceans in 1998 and found that the model forced by global observational sea surface temperatures (SST) can reproduce the heavy rainfall over the Yangtze River valley. Precipitation anomalies in China have a relation with climate in different regions of the Pacific and the equatorial Indian Ocean (Guo *et al.* 2011; Liu *et al.* 2014).

Efficiently identifying the inputs for a streamflow forecast model from so many potential variables is a hot problem. Fernando *et al.* (2009) indicated that the task of an input selection algorithm is to determine the strength of the relationship between potential model inputs and outputs. A traditional method to describe the structure of dependence between variables is linear relation, it is based on the multivariate normal distribution and tends to focus on the degree of dependence, and ignores the structure of dependence. Two important measures of dependence, known as Kendall's tau and Spearman's rho, provide perhaps the best alternatives to the linear correlation coefficient as a measure of dependence for non-Gaussian distributions, for which the linear correlation coefficient is inappropriate and often misleading. The disadvantage of rank-based correlation coefficient is that there is a loss of information when the data are converted to ranks; if the data are normally distributed, it is less powerful than the Pearson correlation coefficient (Gauthier 2001). An alternative method is mutual information, a measure of the amount of information that one random variable contains or explains about another random variable. It can be used to indicate the dependence or independence between variables. However, the MI method is not directly able to deal with the issue of redundant inputs (Bowden *et al.* 2005a, 2005b). Then a copula-entropy (CE) method was proposed to calculate mutual information and partial mutual information, which characterizes the dependence between potential model input and output variables directly instead of calculating the marginal and joint probability distributions (Chen *et al.* 2014a, 2014b). In our study, an all relevant feature selection wrapper algorithm (Kursa *et al.* 2010) is used to determine the final model inputs. The method performs a top-down search for relevant features by comparing the importance of the original attribute with the importance achievable at random, estimated using their permuted copies, and progressively eliminating irrelevant features to stabilise that test.

The remainder of this paper is organized as follows: the next section presents a brief description of theory and methods involved in this study. The study area and utilized data are then described, and then a case study with results and discussion are revealed, and finally summary of the conclusions from this study are presented.

METHODS

A seasonal streamflow forecasts using mixture-kernel GPR and a top-down relevant features search method is constructed in this study. First, the Box-Cox data transformation is employed to make skewed seasonal streamflow and potential climate variables follow a normal distribution. Then a relevant feature-selection wrapper algorithm is applied to determine the model inputs. GPR using mixture-kernel is selected as driven model for its ability of probabilistic streamflow forecasting. Finally, the proposed model performance is examined with five error evaluation indexes.

In the following, methods of Box-Cox data transformation, relevant feature-selection wrapper algorithm, GPR and forecast verification are introduced.

Box-Cox transformation

Hydrologic time series have a varying mean and changing variance over time (Chen *et al.* 2017). The distribution of monthly and seasonal streamflow tends to be highly skewed. Climate indices can be skewed too, although usually only moderately. In GPR streamflow forecasting model, the Box-Cox data transformation (George & Foster 2000) is employed to make skewed variables follow a normal distribution. To ease statistical inference, data transformation parameters are estimated using a maximum likelihood method. The one-parameter Box-Cox transformation is applied to streamflow data and climate indices.

$$z = \begin{cases} \dfrac{(y+1)^{\lambda} - 1}{\lambda} & y \geq 0, \lambda \neq 0 \\ \log(y+1) & y \geq 0, \lambda = 0 \\ -\{(-y+1)^{2-\lambda} - 1\}/(2-\lambda) & y < 0, \lambda \neq 2 \\ \log(-\lambda+1) & y < 0, \lambda = 2 \end{cases} \tag{1}$$

where y and z are the original and transformed variables, respectively, and λ is transformation parameter.

Relevant feature-selection wrapper algorithm

In the implementation of machine learning, determination of immediately set predictors influences model accuracy. The greater the number of features, the longer the time required to train the model, and redundant dependencies between features also easily lead to the reduction of generalization ability. Therefore, it is most useful to eliminate redundant features, keep important features and finally confirm the appropriate features subset.

The importance of each feature can be estimated by building a model. Some methods, such as decision trees, have a built-in mechanism to report on variable importance. Boruta is an all relevant feature selection wrapper algorithm, capable of working with Random Forest (Kursa *et al.* 2010). It can be used to build many models with different subsets of a dataset and identify those attributes that are and are not required to build an accurate model. The algorithm is relatively quick, can usually be run without tuning of parameters and gives a numerical estimate of the feature importance.

The method performs a top-down search for relevant features, which consists of following steps (Kursa & Rudnicki 2010). (1) Extend the information system by adding copies of all variables (the information system is always extended by at least 5 shadow attributes, even if the number of attributes in the original set is lower than 5). (2) Shuffle the added attributes to remove their correlations with the response. (3) Run a random forest classifier on the extended information system and gather the Z scores computed. (4) Find the maximum Z score among shadow attributes (MZSA), and then assign a hit to every attribute that scored better than MZSA. (5) For each attribute with undetermined importance perform a two-sided test of equality with the MZSA. (6) Deem the attributes which have importance significantly lower than MZSA as 'unimportant' and permanently remove them from the information system. (7) Deem the attributes which have importance significantly higher than MZSA as 'important'. (8) Remove all shadow attributes. (9) Repeat the procedure until the importance is assigned for all the attributes, or the algorithm has reached the previously set limit of the random forest runs.

Gaussian process

Gaussian process regression

GPR model is nonparametric kernel-based probabilistic model (Rasmussen & Williams 2012). It is important in the field of machine learning. The key idea of GPR is to assume that the learning sample follows the prior probabilities of the Gaussian process and then calculate the corresponding posterior probability. It is developed based on the Bayesian linear regression model. GPR uses the kernel to define the covariance of a prior distribution over the target functions and uses the observed training data to define a likelihood function. Based on Bayes theorem, a (Gaussian) posterior distribution over target functions is defined, whose mean is used for prediction.

The collection of transformed predictors and predictands form a column vector $\mathbf{z}^T = [z_1, z_2, \cdots z_d]$. $z_i \sim N(\mu_i, \sigma_i)$, d is the dimension of predictors and predictands, the joint Gaussian distribution for the d variables

$$\mathbf{z} \sim N(\boldsymbol{\mu}, \Sigma) \tag{2}$$

$$\boldsymbol{\mu}^T = [\mu_1, \mu_2, \cdots \mu_d] \tag{3}$$

$$\Sigma = diag(\boldsymbol{\sigma}) \times \mathbf{R} \times diag(\boldsymbol{\sigma}) \tag{4}$$

where $diag(\boldsymbol{\sigma})$ is a diagonal matrix from the standard deviation, $\boldsymbol{\sigma}^T = [\sigma_1, \sigma_2, \cdots \sigma_d]$, and $\mathbf{R}$ is the symmetric correlation matrix.

Random process is a Gaussian process when the joint distribution of any random variable follows the Gaussian distribution, the interdependence of multiple transformed predictors and predictands is fully characterised by the mean vector, standard deviation vector, and correlation matrix.

In the Gaussian process regress, the transformed variables $\mathbf{z}^T$ is separated into two sub-vectors $\left[\mathbf{x}^T y^T\right]$, where $\mathbf{x}^T$ and y^T contain the predictors and predictand, respectively. Correspondingly, the mean vector and covariance vector are partitioned as follows:

$$\boldsymbol{\mu}^T = [\boldsymbol{\mu}_\mathbf{x}^T \, \boldsymbol{\mu}_y^T] \tag{5}$$

$$\Sigma = \begin{bmatrix} \Sigma(\mathbf{x}, \mathbf{x}) & \Sigma(\mathbf{x}, y) \\ \Sigma(y, \mathbf{x}) & \Sigma(y, y) \end{bmatrix} \tag{6}$$

In the Bayesian analysis of the standard linear regression model with Gaussian noise $\varepsilon \sim N(0, \sigma^2)$

$$y = \mathbf{x}^T \mathbf{w} + \varepsilon \tag{7}$$

In the condition of known training set input $\mathbf{x}$ and output y, given a new input vector $\mathbf{x}_*$, the posterior distribution of the predicted value y_* is deduced. The Gaussian posterior

$$p(\mathrm{y}_*|\mathbf{x}_*, \mathbf{x}, y) = N\left(\frac{1}{\sigma^2} \mathbf{x}_*{}^T A^{-1} \mathbf{x} y, \mathbf{x}_*{}^T A^{-1} \mathbf{x}_*\right) \tag{8}$$

where $A = \sigma^{-2} X X^T + \Sigma_p^{-1}$, Σ_p is covariance matrix of $\mathbf{w}$.

Bayesian linear model suffers from limited expressiveness. A very simple idea to overcome this problem is to project the inputs into a high-dimensional space using a set of basis feature space functions and then apply the linear model in this space instead of directly on the inputs themselves. Thus, introduce the function $\phi(x)$ to project the scalar input x into some high-dimensional feature space to implement polynomial regression.

The Gaussian posterior can be written

$$p(\mathrm{y}_*|\mathbf{x}_*, \mathbf{x}, y) = N\left(\frac{1}{\sigma^2} \Phi(\mathbf{x}_*)^T A^{-1} \Phi y, \Phi(\mathbf{x}_*)^T A^{-1} \Phi(\mathbf{x}_*)\right) \tag{9}$$

where $A = \sigma^{-2} \Phi \Phi^T + \Sigma_p^{-1}$

To make predictions using this equation, we need to invert the A matrix of size $N \times N$ which may not be convenient if N, the dimension of the feature space, is large. However, we can rewrite the equation in the following way.

$$\begin{aligned} p(\mathrm{y}_*|\mathbf{x}_*, \mathbf{x}, y) = N(&\Phi_*{}^T \Sigma_p \Phi (K + \sigma^2 I_n)^{-1} y, \Phi_*^T \Sigma_p \Phi_* \\ &- \Phi_*^T \Sigma_p \Phi (K + \sigma^2 I_n)^{-1} \Phi^T \Sigma_p \Phi_*) \end{aligned} \tag{10}$$

where $\Phi(\mathbf{x}_*) = \Phi_*$, $K = \Phi^T \Sigma_p \Phi$, K is the kernel function.

$$\begin{aligned} p(\mathrm{y}_*|\mathbf{x}_*, \mathbf{x}, y) = N(&K(\mathbf{x}_*, \mathbf{x})(K + \sigma^2 I_n)^{-1} y, K(\mathbf{x}_*, \mathbf{x}_*) \\ &- K(\mathbf{x}_*, \mathbf{x})(K + \sigma^2 I_n)^{-1} K(\mathbf{x}, \mathbf{x}_*)) \end{aligned} \tag{11}$$

Gaussian process can be derived from its mean and covariance matrices (Williams & Rasmussen 2005). Assuming that the training samples obey a normal distribution with a mean of zero, a unique GPR model can be obtained by knowing the kernel function K. The kernel function and the related parameters determine the performance of the GPR model.

Mixture-kernel function

Covariance is a positive definite symmetric matrix, so the covariance function is equivalent to the kernel function. The kernel function and related parameters determine the fundamental performance of the constructed Gaussian process model.

For streamflow time series, several features are immediately apparent: a long-term trend, a pronounced seasonal variation and some smaller irregularities. In the following, we combine covariance function which takes care of these individual properties. To model the long-term smooth trend we use a squared exponential smooth trend covariance term,

$$k_1(x, x') = \exp\left(-\frac{(x-x')^2}{2\theta_1^2}\right) \tag{12}$$

use the periodic covariance function with a period of one year to seasonal component model the seasonal variation

$$k_2(x, x') = \exp\left(-\frac{2\sin^2((x-x')/2)}{\theta_2^2}\right) \tag{13}$$

to model the small irregularities, a rational quadratic term is used

$$k_3(x, x') = \left(1+\frac{(x-x')^2}{2\theta_3\theta_4^2}\right)^{-\theta_3} \tag{14}$$

where θ_1, θ_2, θ_3 is the typical length-scale and θ_4 is the shape parameter determining diffuseness of the length-scales.

Forecast verification

The accuracy of the monthly runoff forecasting is evaluated with five indices: MSLE (Hogue *et al.* 1999; De Vos & Rientjes 2008), M4E (De Vos & Rientjes 2008), R^2, mean relative error (MRE) and root mean square error (RMSE).

MSLE is mean-squared logarithmic error, M4E is mean fourth-power error. The definition of MSLE and M4E are shown as Equations (15) and (16), respectively.

$$MSLE = \frac{1}{N}\sum_{i=1}^{N}(\ln Q_i - \ln \widehat{Q}_i)^2 \tag{15}$$

$$M4E = \frac{1}{N}\sum_{i=1}^{N}(Q_i - \widehat{Q}_i)^4 \tag{16}$$

MSLE and M4E selected in this study are intended to represent different characteristics of the runoff hydrograph. The MSLE function puts more emphasis on low flows due to the logarithmic transformation. The M4E is considered to be an indicator of goodness-of-fit to high flows as larger deviations are given more contributions. MSLE and M4E can measure different aspect of characteristics of the runoff.

Besides, the fitting accuracy on low flows or peak flows is measured by index R^2, MRE and RMSE. The definition of R^2, RME and RMSE are shown as Equations (17)–(19), respectively.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Q_i - \widehat{Q}_i)^2}{\sum_{i=1}^{N}(Q_i - \bar{Q})^2} \tag{17}$$

$$MRE = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|\widehat{Q}_i - Q_i\right|}{Q_i} \tag{18}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Q_i - \widehat{Q}_i)^2} \tag{19}$$

where N is the length of the observed runoff series, Q_i denotes the ith observed runoff, $\widehat{Q}_i$ denotes the ith simulated runoff, $\bar{Q}$ denotes the average of observed runoff.

Study area and data

GPR model is assessed by streamflow forecast on Jinsha River catchment. The study area is located in the upper reaches of Yangtze River, covers an area of 502,000 km^2, with annual average streamflow about 4,750 m^3/s. The largest reservoir on the Jinsha River is Xiangjiaba Reservoir. It is a multi-objective reservoir that produces electrical power and provides irrigation water and domestic water for Sichuan province. From July to September, it also undertakes the task of flood control. The location of Jinsha River and Xiangjiaba Reservoir is shown in Figure 1. The digital elevation model (DEM) maps are obtained through USGS website and GIS software is utilized to develop the watershed boundary. Here we will focus on the inflow prediction of Xiangjiaba Reservoir.

Streamflow and NINO datasets for the period 1961–2013 are used in this study. They are sourced as follows: quality-controlled and partially in-filled streamflow data

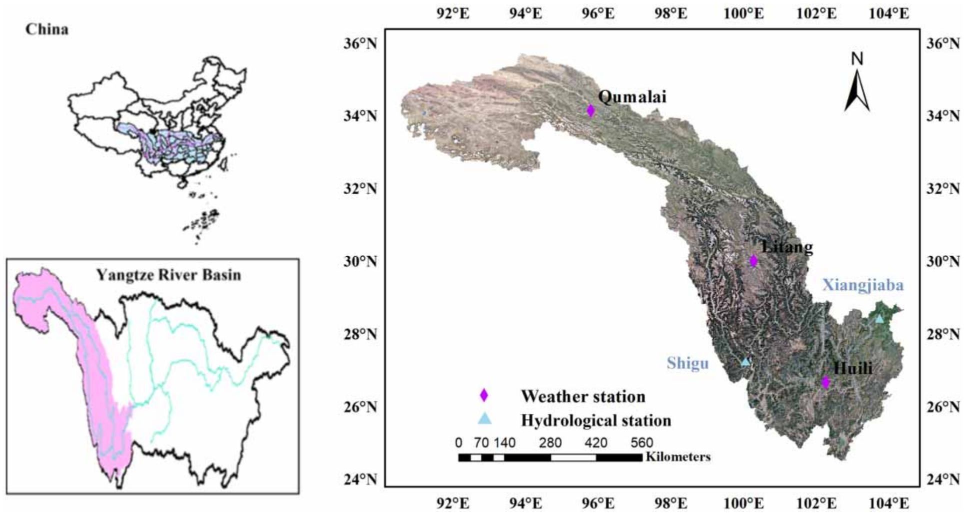

Figure 1 | The location of Jinsha River and Xiangjiaba Reservoir.

measured at Xiangjiaba gauges are provided by the Yangtze River Waterway Bureau; monthly streamflow is calculated from daily streamflow; the SST datasets are described in Table 1, containing east central tropical pacific SST (Nino 3.4), western hemisphere warm pool (WHWP), pacific warm pool, tropical pacific SST EOF, central tropical pacific SST (Nino 4), extreme eastern tropical pacific SST (Nino 1 + 2), Oceanic Niño Index (ONI), and El Niño evolution (TNI); and the SST datasets are downloaded from the National Oceanic and Atmospheric Administration website (http://www1.ncdc.noaa.gov/pub/data/cmb/ersst/v4/netcdf/).

Table 1 | Description of the SST datasets used in this study

Number	Alternative factor	Description
1	Q	Streamflow of Xiangjiaba
2	Nino 3.4	East Central Tropical Pacific SST (5N–5S, 170–120 W)
3	WHWP	Western Hemisphere warm pool
4	Pacific Warmpool	1st EOF of SST (15S–15N, 60E–170E)
5	Tropical Pacific SST EOF	1st EOF of SST (20N–20S, 120E–60 W)
6	Nino 4	Central Tropical Pacific SST (5N–5S, 160E–150 W)
7	Nino 1 + 2	Extreme Eastern Tropical Pacific SST (0–10S, 90 W–80 W)
8	ONI	Oceanic Nino Index
9	TNI	Indices of El Niño evolution

RESULT ANALYSIS

Characteristics analysis on streamflow

The Jinsha River is located in southwest China and is affected by subtropical monsoon climate. Precipitation is the main source of streamflow and almost concentrated in summer, therefore the seasonal distribution of streamflow is extremely uneven (Ye *et al.* 2016). Before the model building, streamflow characteristics on Xiangjiaba are analyzed. Table 2 gives the values of annual mean streamflow, maximum and minimum monthly streamflow. Maximum 19,488 m^3/s occurred in July 1998, and minimum 1,109 m^3/s occurred in March 1995. Then a Mann-Kendall trend test method (Ye *et al.* 2015) is applied to Xiangjiaba

Table 2 | Characteristics of monthly streamflow time series

Streamflow	Annual mean	Maximum	Minimum	MK Trend
Xiangjiaba	4,630	19,488 (1,998.07)	1,109 (1,995.03)	↑

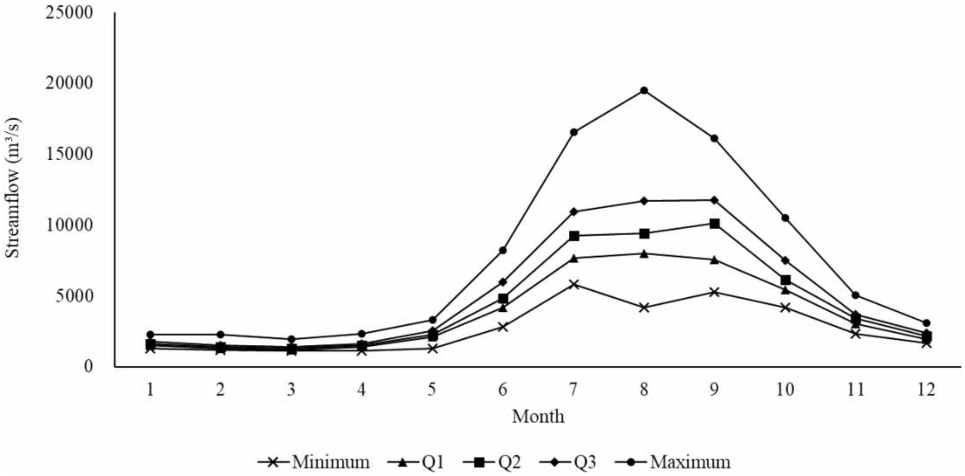

Figure 2 | Quartile graph of monthly streamflow (forecast error increases with streamflow value).

streamflow times series, and a rising trend is tested out. A quartile graph of streamflow is depicted in Figure 2. It can be seen that variance in flood season from June to October is much larger than that in non-flood season. Due to the existence of this heteroscedasticity, we adopt the Box-Cox transformation (Box & Cox 1964; George & Foster 2000) to normalize prediction variable errors and stabilize their variance in the transformed space.

Forecasting factor

Analysis of the relevance and importance of predictors is an extreme step in streamflow forecasting. Different predictor sets mean the difference between mediocre performance with long training times and great performance with short training times.

The relevant feature-selection wrapper algorithm *Boruta* iteratively compares the importance of 108 predictors with weights, or importances, of shadow predictors, created by shuffling the original ones. Eighty-three predictors with significantly worse importance than shadow ones are consecutively rejected or tentative. On the other hand, 25 predictors are significantly better than shadows and are admitted to be confirmed important as shown in Figure 3 and Table 3. The importance analysis indicated that the streamflow and Nino $1+2$ have stronger correlations than the other predictors. The unimportant predictors are also plotted in Figure 4.

Finally, we compared the performance of different sizes of predictor sets using the training data, and found that the set constituted by the top 12 importance predictors gives almost comparable results. All 12 predictors selected in this study are streamflow at 12, 6, 5, 1, 11, 7 months ahead, Nino $1+2$ at 11, 5 months ahead, streamflow at 8 months ahead, Nino $1+2$ at 12, 10 months ahead and streamflow at 4 months ahead, with the sequence of importance.

GPR runoff forecast

In this study, the GPR modelling approach is applied to producing monthly sub-seasonal forecasts. Subsequently, the relationship between transformed predictors and predictands is modelled using a multivariate Box-Cox transformed normal distribution. The kernel used is composed of a squared exponential kernel, periodic kernel and a rational quadratic term. Each term reflects different properties of the dataset: squared exponential kernel explains a long term, smooth rising trend; periodic covariance function with a period of one year to seasonal component model the seasonal variation; a rational quadratic term is used to model the small irregularities; and the hyperparameter value of the kernel function has a great influence on the learning and prediction. In this study, the hyperparameter is optimized using gradient ascent on the log-marginal-likelihood.

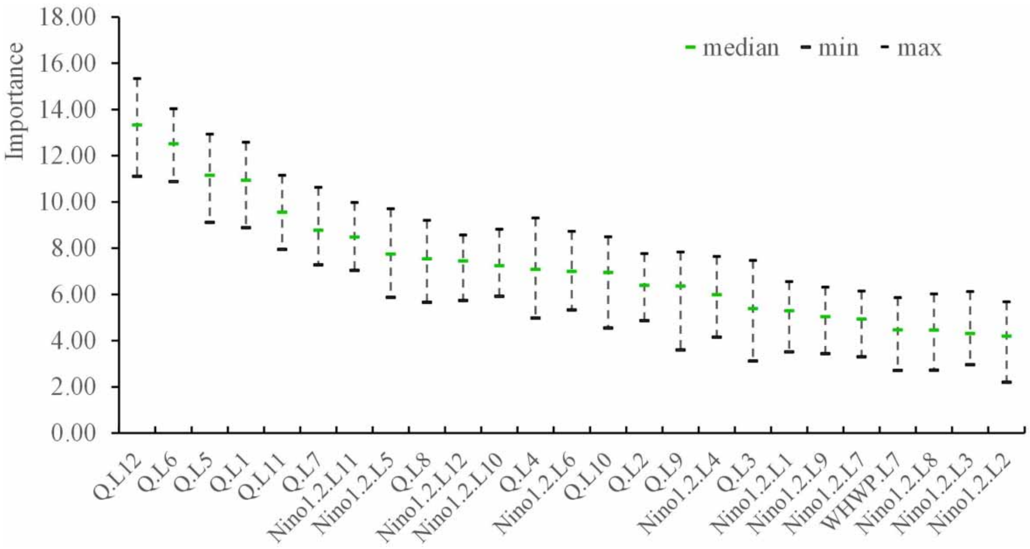

Figure 3 | Predictors confirmed as important by *Boruta* feature selection algorithm.

Table 3 | Important values of top 25 predictors by Boruta feature selection algorithm

No.	Importance	Predictor	No.	Importance	Predictor
1	13.34	Q_{t-12}	14	6.97	Q_{t-10}
2	12.52	Q_{t-6}	15	6.40	Q_{t-2}
3	11.16	Q_{t-5}	16	6.37	Q_{t-9}
4	10.95	Q_{t-1}	17	6.00	Nino $1+2_{t-4}$
5	9.57	Q_{t-11}	18	5.40	Q_{t-3}
6	8.79	Q_{t-7}	19	5.31	Nino $1+2_{t-1}$
7	8.49	Nino $1+2_{t-11}$	20	5.05	Nino $1+2_{t-9}$
8	7.75	Nino $1+2_{t-5}$	21	4.94	Nino $1+2_{t-7}$
9	7.55	Q_{t-8}	22	4.48	$WHWP_{t-7}$
10	7.46	Nino $1+2_{t-12}$	23	4.46	Nino $1+2_{t-8}$
11	7.26	Nino $1+2_{t-10}$	24	4.33	Nino $1+2_{t-3}$
12	7.10	Q_{t-4}	25	4.21	Nino $1+2_{t-2}$
13	7.01	Nino $1+2_{t-6}$			

Monthly scale dataset containing variables of 12 predictors and one predictand in the range 1961–1998 are used for model training, and remaining datasets during 1999–2008 are used for model testing. Four-fold cross validation is implemented for training. Figures 5 and 6 present the mixture-kernel GPR prediction details of the hydrograph in the training and testing periods, respectively. The observed flow is shown in a line, and corresponding prediction mean are shown as dots. As mentioned above, mixture-kernel GPR has a capacity to provide prediction interval as supplementary information when considering prediction uncertainty. From a different angle, we observe that the GP model tends to perform better during the dry season, exhibiting more persistent than erratic flow regimes during hot humid season. For most cases, the confidence envelope obtained by GPR captures streamflow variations adequately during the testing period, except for extreme flash flooding events. The variability of erratic regimes is much more significant than that of the persistent regimes and, thus, is less predictable. The uncertainty bounds shown in Figures 5 and 6 are asymmetric with respect to the average values. This asymmetric phenomenon can also be observed in Sun *et al.* (2014). The reason is that, before training, the streamflow data were normalized using Box-Cox transform and all variables were linearly scaled to the interval $[-1,1]$. After testing, the results were transformed back to the original input space to calculate the performance metrics.

Figures 7 and 8 present the mixture-kernel GPR prediction errors of the hydrograph in the training and testing periods, respectively. Satisfactory performance of GPR streamflow forecasting can also be seen in these two plots.

In order to test the performance of mixture-kernel GPR streamflow forecast model, in this work, we propose to use a highly efficient ANN model, i.e., the generalized

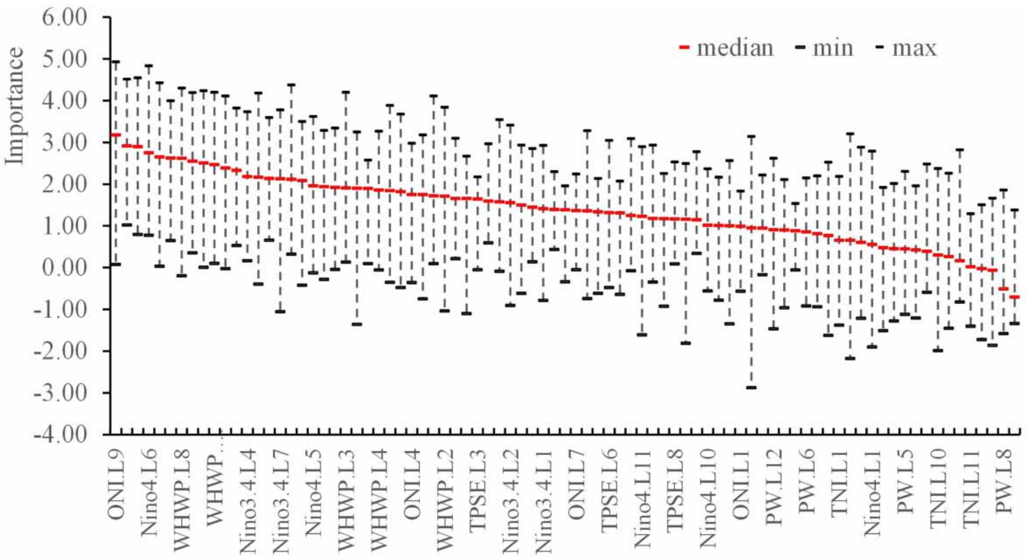

Figure 4 | Predictors confirmed as unimportant by *Boruta* feature selection algorithm.

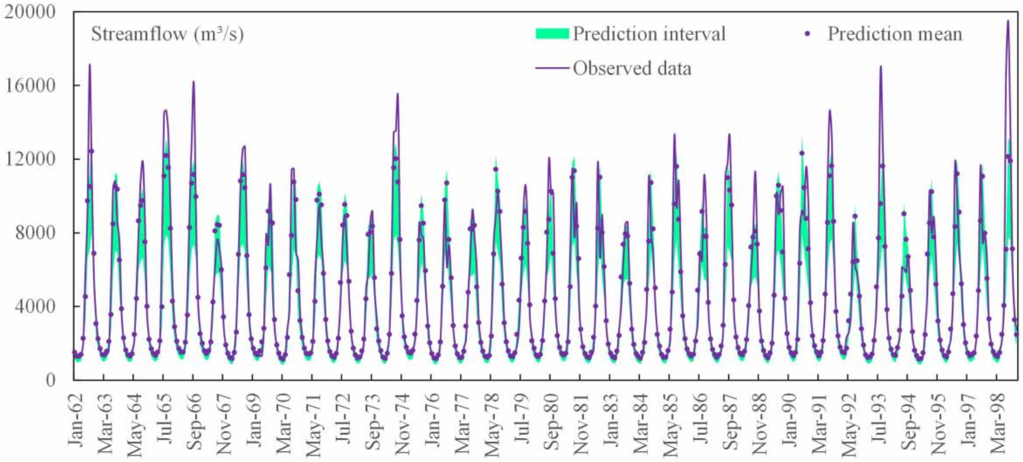

Figure 5 | Mixture-kernel GPR prediction details of the hydrograph in training period.

regression neural network (GRNN), to predict the Xiangjiaba streamflow. The GRNN was proposed after the BP model, and is a kind of RBF ANN. Compared with the BP network, GRNN has good performance for local approximation and does not fall into the local minimum. No iteration is required in its training, and only one parameter needs to be adjusted, i.e. the smoothing factor. For GRNN forecast model, the predictors and datasets partition displays no difference to GPR, and four-fold cross validation is also implemented for training. Figures 9 and 10 display the GRNN prediction details of the hydrograph, and Figures 11 and 12 present the GPR prediction errors of the hydrograph in training and testing periods, respectively. The graphical comparison between GPR and GRNN outputs show that the GPR underestimate the high values, GPR is not as good as GRNN for erratic high flow forecasting as GPR rely on unskewed Gaussian distribution, but overestimation is also a problem for GRNN, which may impact the error statistic values of performance evaluation. We also observe that the GPR model tends to perform better for low and medium flow. Additionally, a single GRNN model only gives prediction dots and cannot

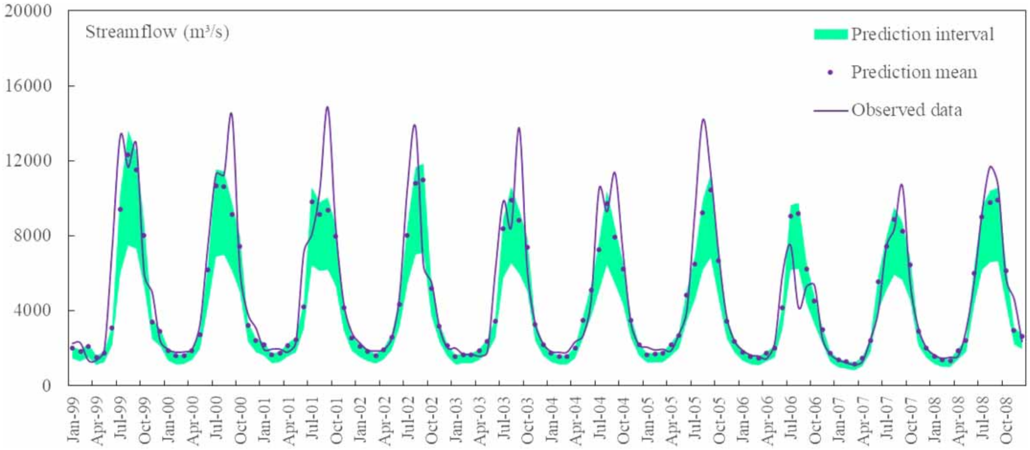

Figure 6 | Mixture-kernel GPR prediction details of the hydrograph in testing period.

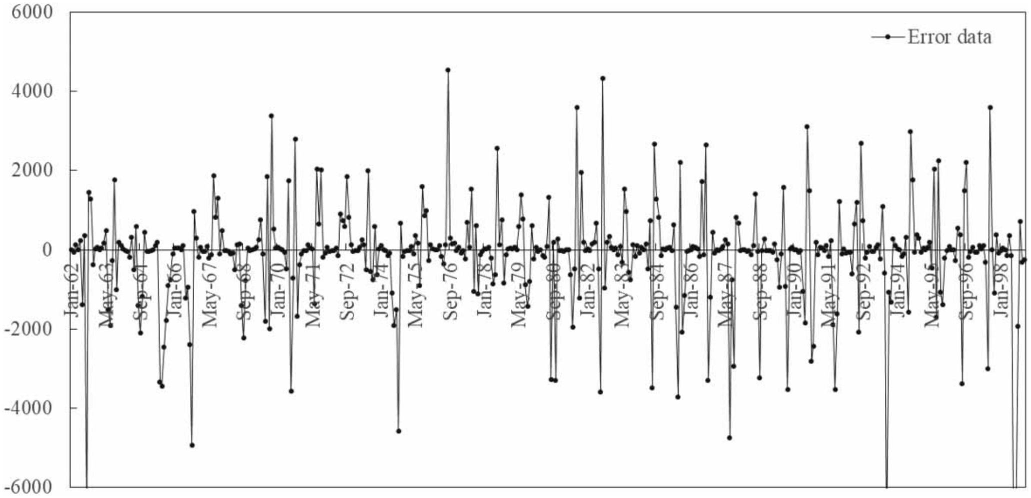

Figure 7 | Mixture-kernel GPR prediction error of the hydrograph in training period.

provide prediction interval. A detail evaluation based on five metrics is given in the following section.

Performance analysis

R^2, RMSE, M4E, MSLE and MRE are used to assess the performance of GPR and GRNN. R^2 measures the degree of coincidence between the streamflow prediction process and the observed process. RMSE focuses on the deviation between the observed value and the predicted value, the formula forms of MSLE and M4E are similar to RMSE, while the MSLE puts more emphasis on low flows due to the logarithmic transformation. The M4E is considered as an indicator of goodness-of-fit for high flows as larger deviations are given more contributions. MRE is average relative error, which presents the error level of forecasting. Five index values for GPR and GRNN forecasting model are given in Table 4. It can be seen that the statistical value of GPR is better than that of GRNN when considering the degree of coincidence, deviation between the observed value and the predicted value, the error level of forecasting, and relatively more emphasis on low flows or high flows.

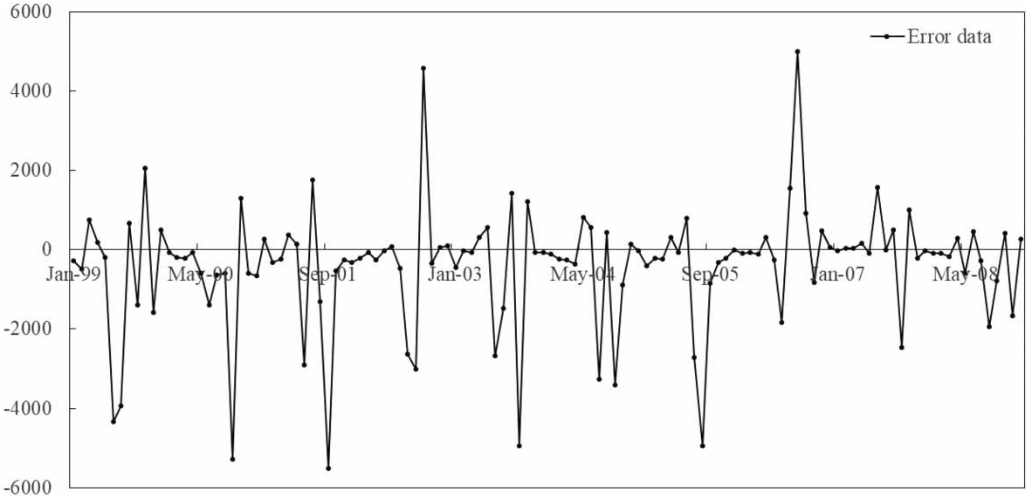

Figure 8 | GPR prediction error of the hydrograph in testing period.

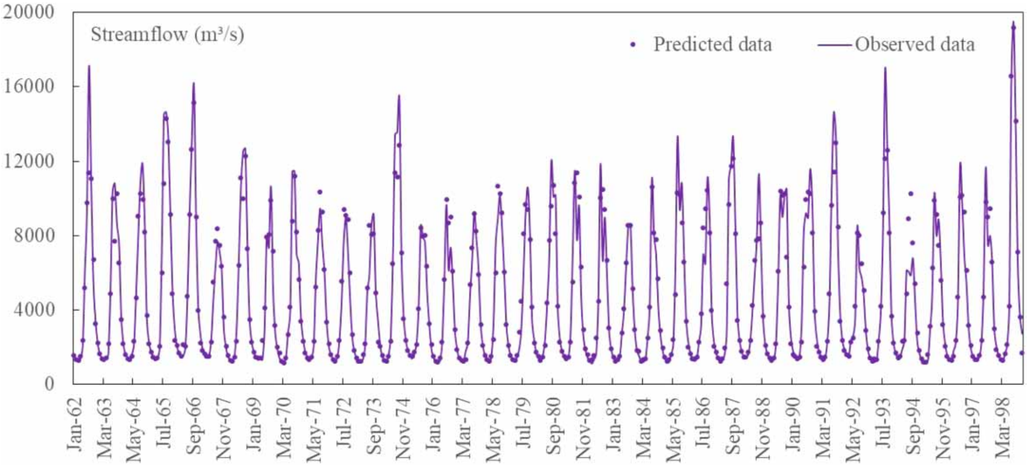

Figure 9 | GRNN prediction details of the hydrograph in training period.

If we only focus on extreme flows, GPR is not as good as GRNN.

CONCLUSIONS

Streamflow simulation using machine learning has been developed and used for different drainage basins in recent decades. In this paper, an emerging non-parametric kernel-based probabilistic model, GPR, is proposed for the streamflow prediction of Jinsha River. The method utilizes a compound kernel function, squared exponential kernel, periodic kernel and a rational quadratic term to reflect different properties of the dataset. A top-down relevant feature-selection wrapper algorithm by random forest is used to determine the final model inputs. In this instance, because hydrologic time series datasets contains trends and seasonality which may result in a varying mean and changing variance over time, a Box-Cox transformation is used to mitigate this kind of heteroscedasticity. The accuracy and presentation of GPR is assessed by comparison with GRNN.

Five indictors selected in this study are intended to represent different characteristics of the runoff hydrograph. The result is the performance of GPR is better than that of

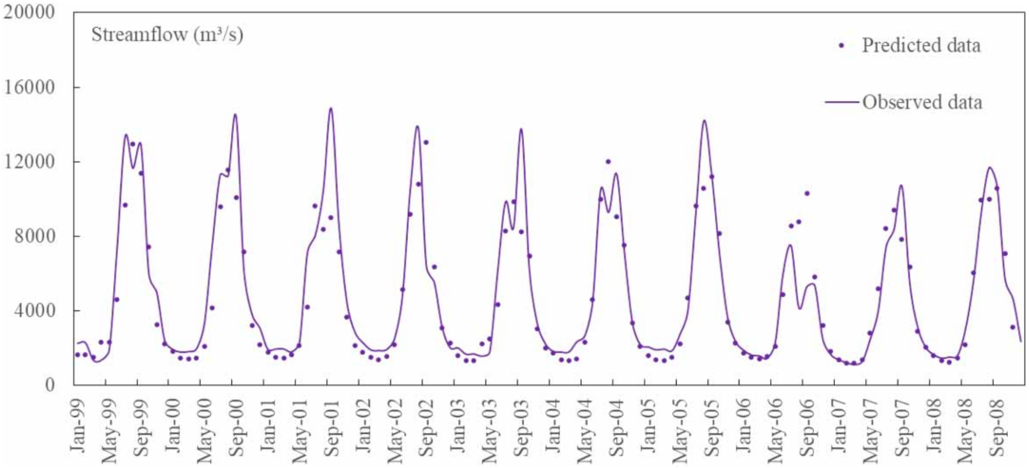

Figure 10 | GRNN prediction details of the hydrograph in testing period.

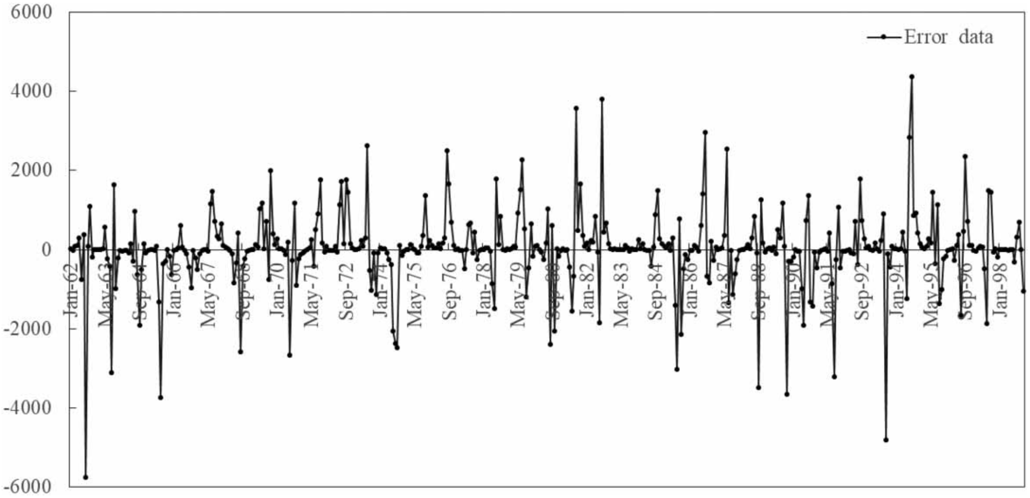

Figure 11 | GRNN prediction error of the hydrograph in training period.

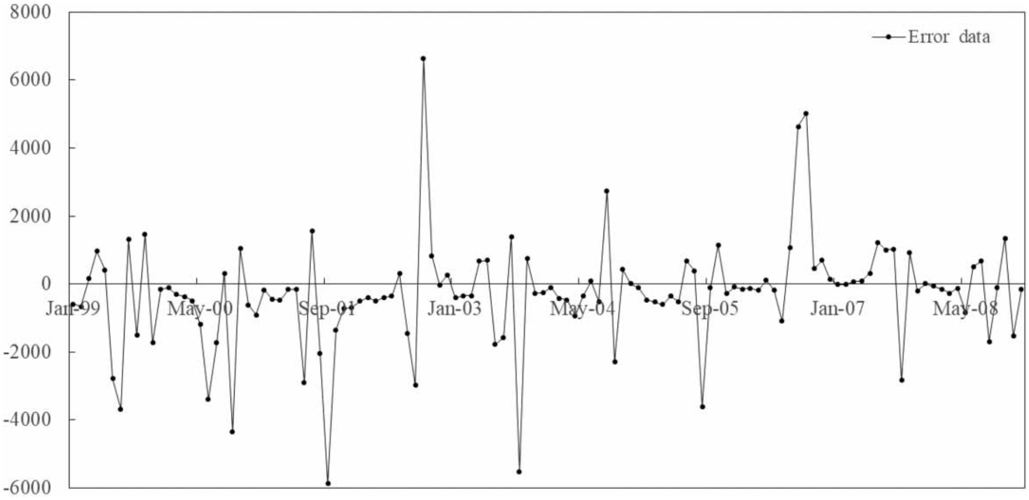

Figure 12 | GRNN prediction error of the hydrograph in testing period.

Table 4 | Values of five error indexs for GPR and GRNN forecasting model

	R^2	RMSE	M4E	MSLE	MRE
GPR	0.84	1,599	4.38×10^{13}	0.052	16.20%
GRNN	0.82	1,638	5.4×10^{13}	0.065	19.70%

GRNN when considering the degree of coincidence, deviation between the observed value and the predicted value, the error level of forecasting, and relatively more emphasis on low flows or high flows. If we only focus on extreme flows, GPR is not as good as GRNN. GPR learns a generative, probabilistic model of the prediction target and can thus provide meaningful confidence intervals and posterior samples along with the predictions while GRNN only provides predictions. The relevant feature-selection wrapper algorithm iteratively compares importances of 108 predictors. Eighty-three predictors, with lower importance than shadow ones, are consecutively rejected or tentative. On the other hand, 25 predictors which have significantly higher importance than shadows are admitted to be confirmed important. The importance analysis indicated that the streamflow and Nino $1+2$ have stronger correlations than the other predictors. We compared the performance of different sizes of predictor sets, and found that the set constituted by the top 12 importance predictors gives best results. All 12 predictors selected in this study are streamflow at 12, 6, 5, 1, 11, 7 months ahead, Nino $1+2$ at 11, 5 months ahead, streamflow at 8 months ahead, Nino $1+2$ at 12, 10 months ahead and streamflow at 4 months ahead, with the sequence of importance.

REFERENCES

Adamowski, J., Chan, H. F., Prasher, S. O., Ozga-Zielinski, B. & Sliusarieva, A. 2012 Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. *Water Resources Research* **48**, 273–279.

Beven, K. 2006 A manifesto for the equifinality thesis. *Journal of Hydrology* **320**, 18–36.

Bowden, G. J., Dandy, G. C. & Maier, H. R. 2005a Input determination for neural network models in water resources applications. Part 1 – Background and methodology. *Journal of Hydrology* **301**, 75–92.

Bowden, G. J., Maier, H. R. & Dandy, G. C. 2005b Input determination for neural network models in water resources applications. Part 2. Case study: forecasting salinity in a river. *Journal of Hydrology* **301**, 93–107.

Box, G. E. P. & Cox, D. R. 1964 An analysis of transformations. *Journal of the Royal Statistical Society* **26**, 211–252.

Chen, L. & Singh, V. 2017 Generalized beta distribution of the second kind for flood frequency analysis. *Entropy* **19**, 254.

Chen, L., Singh, V. P., Guo, S., Zhou, J. & Ye, L. 2014a Copula entropy coupled with artificial neural network for rainfall–runoff simulation. *Stochastic Environmental Research & Risk Assessment* **28**, 1755–1767.

Chen, L., Ye, L., Singh, V., Asce, F., Zhou, J. & Guo, S. 2014b Determination of input for artificial neural networks for flood forecasting using the copula entropy method. *Journal of Hydrologic Engineering* **19**, 217–226.

Chen, L., Singh, V. & Xiong, F. 2017 An entropy-based generalized gamma distribution for flood frequency analysis. *Entropy* **19**, 239.

De Vos, N. J. & Rientjes, T. H. M. 2008 Multiobjective training of artificial neural networks for rainfall-runoff modeling. *Water Resources Research* **44**, 134–143.

El-Shafie, A., Alsulami, H. M., Jahanbani, H. & Najah, A. 2013 Multi-lead ahead prediction model of reference evapotranspiration utilizing ANN with ensemble procedure. *Stochastic Environmental Research and Risk Assessment* **27**, 1423–1440.

Fernando, T., Maier, H. & Dandy, G. 2009 Selection of input variables for data driven models: An average shifted histogram partial mutual information estimator approach. *Journal of Hydrology* **367**, 165–176.

Gauthier, T. D. 2001 Detecting trends using Spearman's rank correlation coefficient. *Environmental Forensics* **2**, 359–362.

George, E. & Foster, D. P. 2000 A new family of power transformations to improve normality or symmetry. *Biometrika* **87**, 954–959.

Guo, Y., Zhao, Y. & Wang, J. 2002 Numerical simulation of the relationships between the 1998 Yangtze river valley floods and SST anomalies. *Advances in Atmospheric Sciences* **19**, 391–404.

Guo, J., Zhou, J., Qin, H., Zou, Q. & Li, Q. 2011 Monthly streamflow forecasting based on improved support vector machine model. *Expert Systems with Applications* **38**, 13073–13081.

Hogue, T. S., Sorooshian, S., Gupta, H., Holz, A. & Braatz, D. 1999 A multistep automatic calibration scheme for river forecasting models. *Journal of Hydrometeorology* **1**, 524–542.

Karimi, S., Shiri, J., Kisi, O. & Xu, T. 2017 Forecasting daily streamflow values: assessing heuristic models. *Hydrology Research* doi:10.2166/nh.2017.111.

Koutsoyiannis, D. 2000 A generalized mathematical framework for stochastic simulation and forecast of hydrologic time series. *Water Resources Research* **36**, 1519–1533.

Koutsoyiannis, D. 2016 Generic and parsimonious stochastic modelling for hydrology and beyond. *International Association of Scientific Hydrology Bulletin* **61**, 225–244.

Koutsoyiannis, D., Yao, H. & Georgakakos, A. 2008 Medium-range flow prediction for the Nile: a comparison of stochastic and deterministic methods/Prévision du débit du Nil à moyen terme: une comparaison de méthodes stochastiques et déterministes. *Hydrological Sciences Journal* **53**, 142–164 (In French).

Kursa, M. B. & Rudnicki, W. R. 2010 Feature selection with the boruta package. *Journal of Statistical Software* **36**, 1–13.

Kursa, M. B., Jankowski, A. & Rudnicki, W. R. 2010 Boruta – a system for feature selection. *Fundamenta Informaticae* **101**, 271–285.

Li, C., Zhou, J., Peng, L. & Wang, C. 2015 Short-term economic environmental hydrothermal scheduling using improved multi-objective gravitational search algorithm. *Energy Conversion & Management* **89**, 127–136.

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. *Journal of Hydrology* **519**, 2822–2831.

Lu, X., Wang, X., Zhang, L., Zhang, T., Yang, C., Song, X. & Yang, Q. 2017 Improving forecasting accuracy of river flow using gene expression programming based on wavelet decomposition and de-noising. *Hydrology Research* doi:10.2166/nh.2017.115.

Markonis, Y. & Koutsoyiannis, D. 2015 Scale-dependence of persistence in precipitation records. *Nature Climate Change* **6**, 399–401.

Montanari, A., Young, G., Savenije, H. H. G., Hughes, D., Wagener, T., Ren, L. L., Koutsoyiannis, D., Cudennec, C., Toth, E. & Grimaldi, S. 2013 'Panta Rhei–Tout s'écoule': Changement hydrologique et sociétal–La Décennie Scientifique 2013–2022 de l'AISH. *Hydrological Sciences Journal* **58**, 1256–1275 (In French).

Papacharalampous, G. A., Tyralis, H. & Koutsoyiannis, D. 2017 Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes. *Journal of Hydrology* **10**, doi:10.20944/preprints201710.0133.v1.

Papacharalampous, G., Tyralis, H. & Koutsoyiannis, D. 2018 Predictability of monthly temperature and precipitation using automatic time series forecasting methods. *Acta Geophysica* **3**, 1–25.

Rasmussen, C. E. & Williams, C. K. I. 2006 *Gaussian Process for Machine Learning*. MIT Press, Boston, Mass pp. 69–106.

Rasmussen, C. & Williams, C. 2012 *Gaussian Processes for Machine Learning*. MIT Press, Boston, Mass pp. 69–106.

Sun, A. Y., Wang, D. & Xu, X., 2014 Monthly streamflow forecasting using Gaussian Process Regression. *Journal of Hydrology* **511**, 72–81.

Tyralis, H. & Koutsoyiannis, D. 2014 A Bayesian statistical model for deriving the predictive distribution of hydroclimatic variables. *Climate Dynamics* **42**, 2867–2883.

Valipour, M., Banihabib, M. E. & Behbahani, S. M. R. 2012 Parameters estimate of autoregressive moving average and autoregressive integrated moving average models and compare their ability for inflow forecasting. *Journal of Mathematics & Statistics* **8**, 330–338.

Williams, C. K. & Rasmussen, C. E. 2005 Gaussian processes for machine learning. *International Journal of Neural Systems* **14**, 69–106.

Wu, C. L., Chau, K. W. & Li, Y. S. 2009 Predicting monthly streamflow using data-driven models coupled with data-preprocessing techniques. *Water Resources Research* **45**, 2263–2289.

Xie, M., Zhou, J., Li, C. & Zhu, S. 2015 Long-term generation scheduling of Xiluodu and Xiangjiaba cascade hydro plants considering monthly streamflow forecasting error. *Energy Conversion and Management* **105** (Supplement C), 368–376.

Ye, L., Zhou, J., Zeng, X. & Tayyab, M. 2015 Hydrological Mann-Kendal multivariate trends analysis in the Upper Yangtze River basin. *Journal of Geoscience & Environment Protection* **3**, 34–39.

Ye, L., Zhou, J., Gupta, H. V., Zhang, H., Zeng, X. & Chen, L. 2016 Efficient estimation of flood forecast prediction intervals via single- and multi-objective versions of the LUBE method. *Hydrological Processes* **30**, 2703–2716.

Yousefi, M., Khoshnevisan, B., Shamshirband, S., Motamedi, S., Md. Nasir, M. H. N., Arif, M. & Ahmad, R. 2015 Support vector regression methodology for prediction of output energy in rice production. *Stochastic Environmental Research and Risk Assessment* **29**, 2115–2126.

Zhu, S., Zhou, J., Ye, L. & Meng, C. 2016 Streamflow estimation by support vector machine coupled with different methods of time series decomposition in the upper reaches of Yangtze River, China. *Environmental Earth Sciences* **75**, 531.

First received 23 January 2018; accepted in revised form 24 April 2018. Available online 8 June 2018