Title

Semi-empirical prediction method for monthly precipitation prediction based on environmental factors and comparison with stochastic and machine learning models

Permalink

https://escholarship.org/uc/item/7g02p9qv

Journal

Hydrological Sciences Journal, 65(11)

ISSN

0262-6667

Authors

Zhang, Huihui
Loáiciga, Hugo A
Ren, Fu
et al.

Publication Date

2020-08-17

DOI

10.1080/02626667.2020.1784901

Peer reviewed

Taylor & Francis
Taylor & Francis Group

Hydrological Sciences Journal

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/thsj20

Semi-empirical prediction method for monthly precipitation prediction based on environmental factors and comparison with stochastic and machine learning models

Huihui Zhang, Hugo A. Loáiciga, Fu Ren, Qingyun Du & Da Ha

To cite this article: Huihui Zhang, Hugo A. Loáiciga, Fu Ren, Qingyun Du & Da Ha (2020) Semi-empirical prediction method for monthly precipitation prediction based on environmental factors and comparison with stochastic and machine learning models, Hydrological Sciences Journal, 65:11, 1928-1942, DOI: 10.1080/02626667.2020.1784901

To link to this article: https://doi.org/10.1080/02626667.2020.1784901

Published online: 07 Jul 2020.

HYDROLOGICAL SCIENCES JOURNAL
2020, VOL. 65, NO. 11, 1928–1942
https://doi.org/10.1080/02626667.2020.1784901

Semi-empirical prediction method for monthly precipitation prediction based on environmental factors and comparison with stochastic and machine learning models

Huihui Zhang[a,b], Hugo A. Loáiciga[b], Fu Ren[a,c,d], Qingyun Du[a,c,d] and Da Ha[b,e]

[a]School of Resource and Environmental Sciences, Wuhan University, Wuhan, China; [b]Department of Geography, University of California, Santa Barbara, California, USA; [c]Key Laboratory of Geographic Information System, Ministry of Education, Wuhan University, Wuhan, China; [d]Key Laboratory of Digital Mapping and Land Information Application Engineering, Ministry of Natural Resources, Wuhan University, Wuhan, China; [e]School of Civil Engineering, Tianjin University, Tianjin, China

ABSTRACT
Precipitation prediction is central in hydrology and water resources planning and management. This paper introduces a semi-empirical predictive model to predict monthly precipitation and compares its predictive skill with those of machine learning (ML) methods. The stochastic method presented herein estimates monthly precipitation with one-step-ahead prediction properties. The ML predictive skill of the algorithms is evaluated by predicting monthly precipitation relying on the statistical association between precipitation and environmental and topographic factors. The semi-empirical predictive model features non-negative matrix factorization (NMF) for investigating the influence of multiple predictor variables on precipitation. The semi-empirical predictive model's parameters are optimized with the hybrid genetic algorithm (GA) and Levenberg-Marquardt algorithm (LM), or GALMA, yielding a validated model with high predictive skill. The methodologies are illustrated with data from Hubei Province, China, which comprise 27 meteorological station datasets from 1988–2017. The empirical results provide valuable insights for developing semi-empirical rainfall prediction models.

ARTICLE HISTORY
Received 28 September 2019
Accepted 9 April 2020

EDITOR
A. Fiori

ASSOCIATE EDITOR
A. Langousis

KEYWORDS
precipitation; prediction; machine learning algorithms; environmental factors; semi-empirical model

1 Introduction

Precipitation is a key meteorological parameter. The identification of suitable models for predicting future precipitation is of primary importance in water resource management, agriculture and flood prevention. Many studies have been conducted on quantitative precipitation predicting using diverse techniques, including application of the autoregressive integrated moving average (ARIMA) model, artificial neural networks (ANNs), support vector regression (SVR) and multiple linear regression analysis (Box *et al.* 2015). Predictive models may be broadly divided into two groups: stochastic models and machine learning models (ML) (Papacharalampous *et al.* 2019).

Among the most popular stochastic models are the fractional Gaussian noise (fGn) and seasonal ARIMA (SARIMA). López-Lambraño *et al.* (2018) employed the Hurst exponent to analyse the persistence of rainfall in a semiarid region of Mexico. Karmakar *et al.* (2019) analysed the long-term memory of monthly and seasonal surface temperature time series in eastern India with the rescaled range analysis (R/S) method. Murthy *et al.* (2018) explored the autocorrelation pattern of rainfall and adopted the SARIMA method to develop a prediction model in northeastern India. Eni and Adeyeye (2015) indicated that the SARIMA model was adequate for predicting rainfall in Warri Town, Nigeria. Shi *et al.* (2015) applied the deep learning long short-term memory (LSTM) model for predicting precipitation. Kumar *et al.* (2019) applied LSTM to predict monthly rainfall and analysed rainfall time series in India.

The successful application of various data-driven models has opened new avenues for the application of machine learning in the field of precipitation prediction. The key advantage of machine learning is learning automatically from the data without resorting to human expertise (LeCun *et al.* 2015). There are various ML models for regression analysis, such as the ANN, regression trees, SVR, Gaussian process regression models (GPR) and others. ANNs may simulate nonlinear systems without any required assumptions in most traditional statistical approaches (Liu *et al.* 2013). Mehdizadeh *et al.* (2018) verified that ANNs have a better performance than the gene-expression programming (GEP) model for estimating rainfall. A review of SVR prediction applications can be found in Sapankevych and Sankar (2009). Bahram *et al.* (2018) reported that regression trees achieved better results compared to the adaptive neuro-fuzzy inference system and the ARIMA model in predicting precipitation. A GPR model was applied to predict solar radiation and was compared with other models by Voyant *et al.* (2017). GPR models are seldom utilized for hydrological process prediction. The applicability of machine learning algorithms in rainfall prediction has been reported by several authors (see, e.g. Mekanik *et al.* 2013, Ramana *et al.* 2013, Abbot and Marohasy 2014; Papacharalampous *et al.* 2018a, 2018b, 2019). Precipitation predicting research often focuses on the comparison between stochastic (mainly ARIMA) and ML methods based on time series of precipitation. Papacharalampous *et al.* (2018a) investigated the multi-step ahead predictability of monthly precipitation by applying

CONTACT Qingyun Du qydu@whu.edu.cn School of Resources and Environmental Sciences, Wuhan University, Wuhan, China; Da Ha Department of Geography, University of California, Santa Barbara, California, USA

seven automatic univariate time series predicting methods to a sample of 1552 monthly precipitation time series. Papacharalampous *et al.* (2019) compared 11 stochastic and 9 ML methods by conducting 12 computational experiments based on simulations. The state-of-the-art approaches have advanced the field of rainfall prediction; yet, predicting precipitation based on ML models with environmental factors that explore the potential statistical association between precipitation and such factors are not customary in precipitation prediction.

Precipitation can be approximated by a linear system, yet, rainfall processes are stochastic in nature and governed by multiple factors (Chinchorkar *et al.* 2012). At the local scale rainfall processes are affected by environmental factors such as terrain characteristics, temperature, humidity and vegetative cover (Pal *et al.* 2019). There is a complex interaction between precipitation and environmental factors. Zhang *et al.* (2003) investigated the feedback effects of vegetative cover on summer precipitation and their results implied that vegetative cover strongly affected summer precipitation in China. Wu *et al.* (2009) proposed that terrain characteristics played an important role in precipitation. However, the direct and indirect effects of environmental factors on precipitation were not quantified. Non-negative matrix factorization (NMF) has been applied to feature extraction from images, the identification of distinct molecular patterns and automatic speech recognition (Lee and Seung 1999, Novak and Mammone 2001, Brunet *et al.* 2004) and plays a significant role in characteristics identification.

A high-precision local model for rainfall prediction is difficult to implement because of resolution challenges that arise when global models are downscaled for local evaluations (Colette *et al.* 2012, Khan *et al.* 2019). Empirical and semi-empirical methods, on the other hand, may be efficient tools for short-term rainfall prediction. Machine learning algorithms are cumbersome and commonly excise a heavy computational burden (Hashemifard *et al.* 2019). Papacharalampous *et al.* (2018a, 2018b) related the performance of stochastic and machine learning models to the size of the solved problem. Milanic *et al.* (1998) compared the performance of a semi-empirical and a neural network model in predicting precipitation rates of TiO_2 particles in an industrial hydrolysis process, determining that the semi-empirical model was the better choice when approximate results were acceptable. Dirks *et al.* (2003) presented a simple semi-empirical model for predicting the effect changes in traffic flow on carbon monoxide concentrations. The building of semi-empirical model mainly suffers from two challenges: the selection of appropriate model factors and the specification of optimization of parameters.

The aim of this work is threefold: first, to assess the accuracies of several stochastic and machine learning algorithms for short-term monthly rainfall prediction with time series and with environmental and topographic factors, respectively; secondly, to explore the functional dependence between monthly precipitation and predictor variables; and thirdly, to propose a semi-empirical precipitation-prediction model that has been benchmarked against multiple ML algorithms. This paper first assesses the accuracies of several stochastic and machine learning algorithms for monthly rainfall prediction in Hubei Province, China, based on univariate predictors (time series) and multiple predictors (environmental and topographic factors), respectively. The performances of the algorithms based on their predictive skills are compared and analysed. Subsequently, NMF is applied for exploring the dominant environmental factors governing precipitation. Lastly, a semi-empirical model for predicting rainfall based on environmental and topographical factors is developed and calibrated based on observation data collected from 27 meteorological stations. The semi-empirical model is benchmarked against multiple ML algorithms. The genetic algorithm/Levenberg-Marquardt algorithm (GALMA) is employed to search for the optimal parameters of the semi-empirical model in this study. The methodology is illustrated with data from Hubei Province, China.

2 Material and methods

2.1 Study area

Hubei Province is centrally located in China (108°21′42″–116°07′50″E, 29°01′53″–33°06′47″N), with an abundance of water and plant resources including dense river networks and developed water systems. Currently, there are over 4000 large and small rivers, totalling channel length of over 60 000 km. Thus, this region is called “the province of the thousand lakes.” Hubei has three diverse topographic zones, including mountain, plain and hill zones, which occupy 56%, 20% and 24% of the total area, respectively. In the northern, eastern and western regions the terrain is high, while in the middle of Hubei Province the terrain features minimal relief. The province is located in the subtropical zone, with a tropical monsoonal climate that is characterized by abundant precipitation, long hours with sunlight and high temperatures for July–August. Its mean annual temperature ranges from 13°C to 18°C and the average annual precipitation ranges between 800 and 1609 mm. Summer (July–August) is the main flooding season of Hubei Province due to the subtropical monsoonal climate and topography. The uneven spatiotemporal distribution of precipitation causes 72% of the annual total precipitation from May–September.

2.2 Data description

2.2.1 Precipitation and humidity data

The precipitation data used in this study include daily rainfall for the period 1988–2017 and monthly rainfall, which was calculated by the accumulation of daily precipitation for the period 2005–2007. The mean monthly relative humidity was calculated from the daily relative humidity from 2005–2007. All data, including their latitude and longitude, were collected at 27 meteorological stations, which cover the entire Hubei Province. The meteorological data came from China’s surface climate daily value (V3.0) dataset supplied by the National Meteorological Information Center. The 27 meteorological stations constitute the 27 test areas, which are displayed in Fig. 1.

2.2.2 Topographic data

A digital elevation model (DEM) contains the basic potential factors affecting precipitation. A 1-km resolution DEM was obtained from the International Scientific & Technical Data

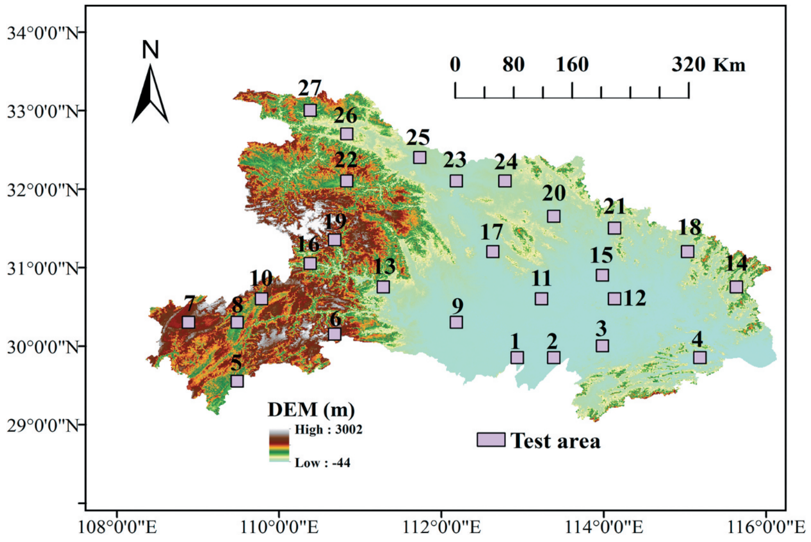

Figure 1. Meteorological stations in Hubei Province, labelled 1 to 27.

Mirror Site, Computer Network Information Center, Chinese Academy of Sciences.[1] Topographic factors, including the slope and altitude, were derived from this DEM. The slope map of Hubei Province is shown in Fig. 2.

2.2.3 Vegetation data

The normalized difference vegetation index (NDVI) is evaluated as one of the potential factors that influence precipitation. The monthly vegetation index L3 dataset was calculated with the Moderate Resolution Imaging Spectroradiometer (MODIS; 1-km resolution) product from the mean daily values every month. The dataset was provided by the International Scientific & Technical Data Mirror Site, Computer Network Information Center of the Chinese Academy of Sciences.[2] The relationships of the NDVI across different months with precipitation indicated that there were differences in the effects of precipitation on the NDVI among the months during the growing season and the response of the NDVI to climate factors lagged (Bao *et al.* 2007, Cao *et al.* 2011, Jedrzejek *et al.* 2013). The "time lag factor' of the vegetation response to precipitation was analysed in the former study. The delayed time is approximately 2 months.

2.2.4 Daytime surface temperature data

The daytime surface temperature (LTD) is considered as another potential factor on rainfall. The LTD data were derived from the MODLT1 T product by taking the mean value every month. The dataset was provided by the International Scientific

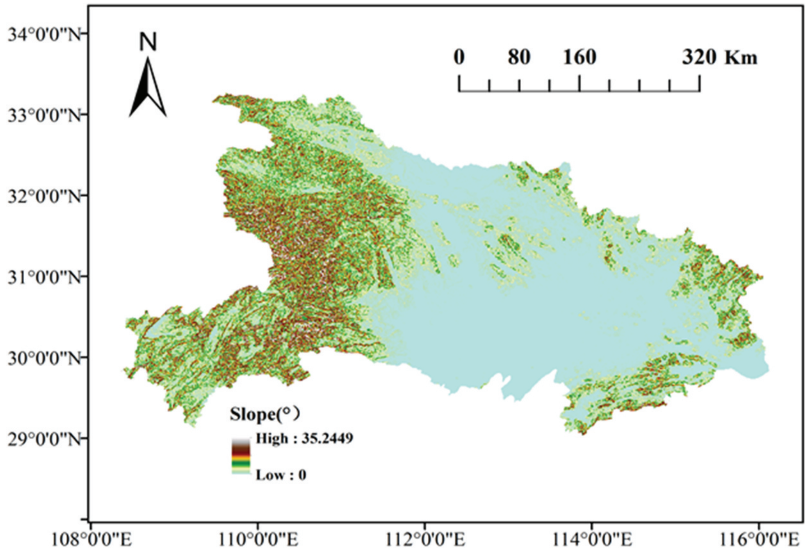

Figure 2. Slope map of Hubei Province.

[1] http://www.gscloud.cn.
[2] http://www.gscloud.cn.

& Technical Data Mirror Site, Computer Network Information Center, Chinese Academy of Sciences[3]; its spatial resolution is 1 km and the time resolution is one month. All these data were projected to the same coordinate system with identical geometric correction, atmospheric correction, sensor correction and cloud processing.

2.2.5 Evaluation parameters

This work relies on the coefficient of determination (R^2), root mean square error (RMSE), mean square error (MSE) and mean absolute error (MAE) to evaluate the predictive skill of several precipitation prediction models. They are calculated as follows:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}\left(X_{pi} - X_{oi}\right)^2}{N}} \tag{1}$$

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(X_{pi} - X_{oi}\right)^2 \tag{2}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|X_{pi} - X_{oi}\right| \tag{3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(X_{oi} - X_{pi}\right)^2}{\sum_{i=1}^{N}\left(X_{oi} - \overline{X_o}\right)^2} \tag{4}$$

where X_{oi} and X_{pi} denote the ith observed and predicted value, respectively ($i = 1, 2, \ldots, N$), where N denotes the total number of observations or predicted data; and $\overline{X_o}$ denotes the mean of the observed data.

2.2.6 Test areas

Six test areas were selected among the 27 meteorological stations as examples (Test Areas 1, 2, 3, 4, 7 and 8; see Fig. 1). The six test areas feature unique morphological characteristics that cover all landform types in Hubei Province, thus demonstrating the generality of this paper's methodology in a wide range of precipitation contexts. The average elevation of Test Area 7 is above 1200 m a.s.l., which is significantly higher than those of the other test areas. Test Area 7 is a mountainous region with pronounced slopes and relatively large NDVI. The average elevation of Test Area 8 is near 500 m a.s.l., with less mountainous features than Test Area 7. The other tests areas encompass plain landforms, with elevations below 200 m a.s.l., and feature mild slopes. The NDVI of Test Area 2 is the lowest among the six test areas measured at a monthly time scale.

2.3 Methodology

2.3.1 Univariate predictor variable prediction methods

The R/S algorithm is applied to analyse the long-term dependence of monthly precipitation. The SARIMA and LSTM algorithms' predictive skill is evaluated respectively based on one-step ahead predicting of monthly precipitation for the period 1988–2017. Prior to the application of the SARIMA and LSTM methods, we divided each precipitation time series into three segments, i.e., the training segment, the testing segment and the validation segment. The training and testing segments are composed of time series for the period 1988–2012. The last five years were chosen as the validation dataset (2013–2017).

2.3.1.1 The R/S method. The Hurst parameter of the fGn can be estimated with the R/S estimator, while the magnitude of the long-range dependence is characterized by the value of the Hurst coefficient, with higher values related to strong long-range dependence (Papacharalampous *et al.* 2018a).

The mathematical basis of the R/S method is described in Beran *et al.* (2013, pp 410–412). It is essential to remove periodic structures of R/S (Mandelbrot and Wallis 1969, Shumway and Stoffer 2017). The second-order IIR notch filter algorithm is herein applied to monthly precipitation to obtain a stationary time series. The reader is referred to Chaparro (2019, PP 639–707) for its definition. The implementation of the filtering algorithm in this work set the notch frequency (w0) equal to 1/12, where 0< w0 < 1. The bandwidth (bw) was set equal to w0 × 10. Lastly, the time series were filtered.

This paper employs the R/S method to explore the magnitude of long-range dependence in long time series of detrended precipitation. All the statistical analyses were programmed with MATLAB R2019a.

2.3.1.2 The SARIMA model. The SARIMA model has been shown to perform better than the simple ARIMA model (Kumar and Lelitha 2015). SARIMA model is a special case of ARIMA models with seasonality. The SARIMA model is applied to analyse time series and forecasting future events in a series. The full model is described in Shumway and Stoffer (2017) and Box *et al.* (2015). The detail procedures of the SARIMA implementation can be found in Pedregal (2019). It is possible to estimate the appropriate values of the autoregressive order p and the moving average order q from the partial autocorrelation function (PACF) and the autocorrelation function (ACF) plots, respectively. Determination of (p, q) is also possible with the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). The optimal parameters correspond to the smallest AIC and BIC and the white noise is tested according to the Ljung-Box statistic (Ljung and Box 1978).

The SARIMA model that captures times series periodicity is herein applied to predict monthly precipitation in this paper. The SARIMA algorithm was implemented by using the Econometrics Toolbox 5.2 of MATLAB R2019a (Chen and Boccelli 2018).

2.3.1.3 The LSTM model. Recurrent neural networks (RNNs) are a type of sequential model applied to predict time series data. The LSTM is a special RNN that captures long-range dependencies and nonlinear dynamics. The LSTM includes one input layer, one or more hidden layers and one output layer. A LSTM cell comprises three interactive neural networks, called the forget gate f_t, input gate i_t and output gate o_t. The principle of LSTM conducted in this study is mainly based on Hochreiter and Schmidhuber (1997). The forget gate (f_t) may drop values

[3] http://www.gscloud.cn.

that are not needed and keep those that are necessary for prediction. The input gate (i_t) renders the new cell state C_t; it is calculated as follows:

$$i_t = \sigma(U_i x_{t-1} + W_i h_{t-1} + b_i) \quad (5)$$

The intermediate memory cell $\widetilde{C}_t$, which is a combination of the input from the last hidden state h_{t-1} and the input x_{t-1}, is calculated as follows:

$$\widetilde{C}_t = \tanh(U_{\tilde{c}} x_{t-1} + W_{\tilde{c}} h_{t-1} + b_{\tilde{c}}) \quad (6)$$

where $\boldsymbol{U}$ and $\boldsymbol{W}$ denote the adjustable weights matrix or learning rates and $\boldsymbol{b}$ denotes the bias vector. Such as $U_{\tilde{c}}$, $W_{\tilde{c}}$ and $b_{\tilde{c}}$ are adjustable weights matrix and vectors of $\widetilde{C}_t$. These matrices or vectors can be optimized in the training of neural networks. tanh denotes the hyperbolic tangent. In the next phase the neuron state is updated according to the following equation:

$$C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C}_t \quad (7)$$

where $\odot$ denotes element-wise multiplication.

Finally, the output gate o_t determines which values are selected by combining o_t with the tanh-modified state C_t having output h_t. The cell state C_t and output h_t are passed to the next time step and go through the forget, input and output gates as outlined in Fig. 3. Iterations are repeated until reaching a stopping criterion.

This paper applies the LSTM model, which has performed well in predicting monthly rainfall. It was implemented with the Deep Learning Toolbox of MATLAB R2019a. The LSTM model's predictive accuracy was herein compared with that of the SARIMA model.

2.3.2 Multivariate predictor variables prediction methods

The machine learning algorithms are applied to predict monthly precipitation based on environmental and topographic factors. The ML methods used in this study are mainly supervised regression learning algorithms. The ML methods are classified into six main categories: back-propagation artificial neural network (BP-ANN); linear regression models; regression trees; SVRs; GPRs; and ensembles of trees, which are listed in Table 3. The implementations of algorithms were mostly made through the MATLAB R2019a Statistics and Machine Learning Toolbox 11.5, which was applied to data mining by Martin *et al.* (2017), while the BP-ANN was programmed with MATLAB 2019 based on Alireza and Peter (2018, p. 20–34). The machine learning algorithms were evaluated based on multivariate predictor variable (LTD, altitude, NDVI, humidity, latitude, longitude, slope and month date) from 2005 to 2007. The first 33 months of the data were applied for training and testing dataset, while the last 3 months of the data served to evaluate the accuracy of the prediction. The primary prediction algorithms are well documented in the literature. The input-output data matrix $\boldsymbol{B}$ is designed as follows:

$$B = [\text{input}(\text{LTD}, \text{altitude}, \text{NDVI}, \text{humidity}, \text{latitude}, \text{longitude}, \text{slope}, \text{month date}), \text{output}(\text{rainfall})]$$

Linear regression models are relatively simple and involve relatively low computational burden for making predictions. However, their low predictive accuracy constrains their uses. Regression trees are easy to interpret, fast for fitting and prediction and low on memory usage. Detailed information about regression trees is available in, for instance, Breiman *et al.* (1984) and Loh (2002). This study sets the default minimum leaf size of the fine tree, medium tree and coarse tree equal to 4, 12 and 36, respectively. The SVRs include the linear SVRs and nonlinear SVRs. The former SVRs is relatively simple but have

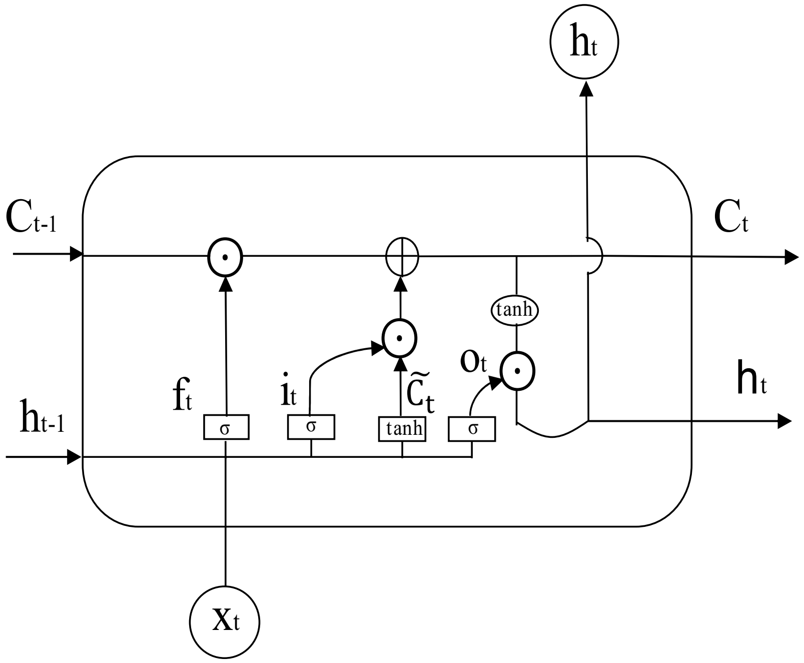

Figure 3. Structure of one LSTM memory block.

low predictive accuracy, while the latter is more complex and accurate. Herein we apply the SVR algorithms by Fan *et al.* (2005, 2006)). GPR models are often highly accurate but can be difficult to interpret. The GPR algorithms feature design and kernel function options that are available in Rasmussen and Williams (2006). Ensemble models combine results from many weak learners into one high-quality ensemble models including boosted trees and bagged trees. Boosted trees are composed of least-squares boosting (LSBoost) and regression tree learners, while bagged trees combine bootstrap aggregating or bagging with regression tree learners (see Breiman 1996, Warmuth *et al.* 2006). The parameters of all the models except BP-ANN are automatically chosen by using hyperparameter optimization. An optimization scheme is applied to seek to minimize the model MSE and return a model with the optimized hyperparameters by testing different combinations of hyperparameter values.

2.3.2.1 BP-ANN. The ANNs are commonly applied to perform large-scale parallel calculations to simulate nonlinear correlation (Catalogna *et al.* 2012). Among the tens of ANN models, the back propagation (BP) network is the most widely used. The basic principle of the algorithm is that after the input signal is presented to the network, the error vector across output units is calculated and back propagated to update the weights (Nourani *et al.* 2011). This study applies the BP-ANN with five layers (one input layer, three hidden layers and one output layer). Each hidden layer includes 6 nodes. The BP-ANN was programmed with MATLAB R2019a. The activation function is a sigmoid function, whose formula is:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{8}$$

The datasets are divided into two parts, the training data and the testing data (90% and 10%, respectively). The prediction precision performance of the BP-ANN training set is evaluated using the testing data for cross validation (Liu *et al.* 2013). The prediction results are evaluated using R^2, RMSE, MSE and MAE.

This paper compares the prediction accuracy of BP-ANNs with those of other ML algorithms. A proposed semi-empirical prediction model is herein identified and benchmarked against the BP-ANN's predictions.

2.3.3 Parameter selection and optimization algorithms

2.3.3.1 NMF. Non-negative matrix factorization was introduced by Paatero and Tapper (1994) and popularized in an article by Lee and Seung (1999). NMF has previously been shown to be a useful decomposition for multivariate data belonging to unsupervised learning. The principle of the NMF algorithm is to find non-negative matrix factors $\boldsymbol{W}$ and $\boldsymbol{H}$ for a given non-negative matrix $\boldsymbol{V}$. NMF captures the data traits by identifying the correlation between data parts and finding the internal interactions among the data. The basic formula is as follows:

$$\boldsymbol{V}_{(F\times N)} \approx \boldsymbol{W}_{(F\times K)}\boldsymbol{H}_{(K\times N)} \tag{9}$$

where $\boldsymbol{V}$ denotes the origin matrix of size $F \times N$. F and N denote the number of rows and columns, respectively. $\boldsymbol{W}$ and $\boldsymbol{H}$ denote the weighting and coefficient matrices, respectively. K is chosen to be smaller than F or N so that $(F + N)\times K < F \times N$, which means $\boldsymbol{W}$ and $\boldsymbol{H}$ are of smaller size than the matrix $\boldsymbol{V}$.

The objective function for finding $\boldsymbol{H}$ and $\boldsymbol{W}$ is written as follows:

$$\min f(\boldsymbol{W}, \boldsymbol{H}) = \frac{1}{2}||\boldsymbol{V} - \boldsymbol{W}\boldsymbol{H}||_F^2, \boldsymbol{W}, \boldsymbol{H} \geq 0 \tag{10}$$

in which $||\cdot||_F^2$ denotes the Frobenius norm.

The optimization of the objective function is made according to the multiplicative iterative principle (Lee and Seung 1999).

This paper employs NMF to detect the sensitive environmental and topographic factors that have a strong statistical association with precipitation to build the semi-empirical prediction model.

2.3.3.2 GALMA. The genetic algorithm/Levenberg-Marquardt algorithm is a hybrid algorithm combining the advantages of the genetic algorithm (GA) and the Levenberg-Marquardt (LM) method to estimate optimal parameters. The GA is first applied to find a suboptimal solution from a global search space. Subsequently, the solution obtained with the GA serves as an initial search point to launch the LM algorithm to achieve a near-global solution that avoids local, suboptimal, solutions. The new parameter optimization algorithm proposed by this study is based on the work of Zheng *et al.* (2019). The optimal parameters of the semi-empirical prediction model are estimated by GALMA in this work. The GALMA was programmed with MATLAB R2019a based on the flow diagram of GALMA (Fig. 4).

2.3.3.2.1 Objective function and termination criterion. Equation (11) is the objective function which measures the agreement between the observed and estimated data:

$$\text{SEE} = \sqrt{\frac{1}{v}\sum\nolimits_{i=1}^{n} e_i^2} \tag{11}$$

where SEE is the standard error of estimates; e_i denotes the ith difference between the value of the observed precipitation and the estimated precipitation; v represents the degrees of freedom, which equals the number of monthly precipitation records minus the number of estimated parameters.

The MSE is relied upon to evaluate the difference between the observed and estimated precipitations, in which case is applied as the suitable termination criterion to terminate the calculation during the computational process (Jha *et al.* 2006). The formula is shown in Equation (2).

In the optimization search the parameter estimation search process is terminated when the MSE equals or is less than the pre-set threshold, or when a maximum number of iterations is reached.

2.3.3.2.2 GA algorithm. This work adopts the GA for searching a suboptimal solution as the first step. The initial population of solutions is generated randomly. The crossover rate P_c and P_m equal to 0.8 and 0.005, respectively (Reed *et al.* 2000). The number of populations of solutions in each GA iteration is set equal to 200 (Samuel and Jha 2003).

2.3.3.2.3 LM algorithm. The LM algorithm is a modified Gauss-Newton optimization approach in which the parameter

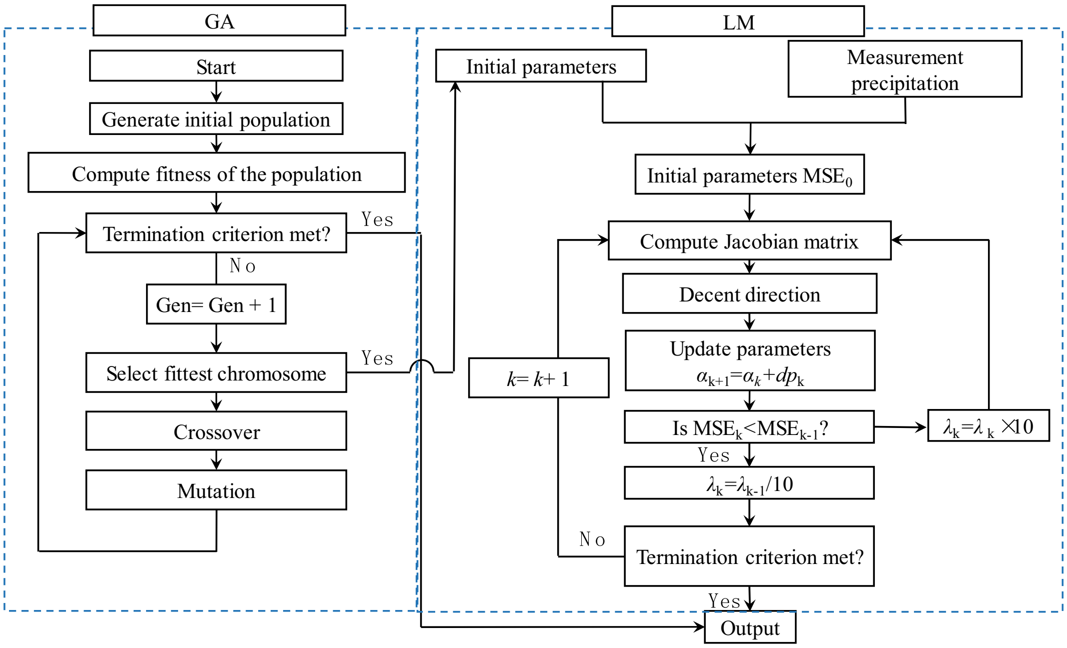

Figure 4. Flow diagram of the GALMA.

set is updated iteratively. The parameters P_1, P_2, P_3, P_4, P_5, and P_6 are considered time-invariant and independent of each other. The termination criterion in this study adopts a maximum number of 200 iterations of the LM algorithm. An initial damping parameter λ_0 is set equal to 10^{-3} (Madsen *et al.* 2004).

The elements of parameter vector $\boldsymbol{\alpha}$ are the six parameters:

$$\alpha^{\mathrm{T}} = [P_1, P_2, P_3, P_4, P_5, P_6]$$

The value of the parameter vector in the kth iteration of the LM algorithm is denoted by $\boldsymbol{\alpha_k}$.

The function $\boldsymbol{g}$ $(\boldsymbol{x}, \boldsymbol{\alpha})$ is employed to calculate monthly precipitation as a function of a vector of independent variables $\boldsymbol{x}$ (say, LTD, altitude, humidity, NDVI) and a vector of parameters. The Jacobian matrix $\boldsymbol{J_k}$ in the kth algorithmic LM iteration reduces to a 6-D vector of derivatives of Equation (12) with respect to the precipitation parameters evaluated in the kth iteration-with $g_k = \boldsymbol{g}$ $(\boldsymbol{x_k}, \boldsymbol{\alpha_k})$:

$$J_k = \left[\frac{\partial g_k}{\partial P_1}, \frac{\partial g_k}{\partial P_2}, \frac{\partial g_k}{\partial P_3}, \frac{\partial g_k}{\partial P_4}, \frac{\partial g_k}{\partial P_5}, \frac{\partial g_k}{\partial P_6}\right]^{\mathrm{T}} \tag{12}$$

where $\frac{\partial g_k}{\partial P_1}, \frac{\partial g_k}{\partial P_2}, \frac{\partial g_k}{\partial P_3}, \frac{\partial g_k}{\partial P_4}, \frac{\partial g_k}{\partial P_5}, \frac{\partial g_k}{\partial P_6}$ can be approximated with the difference equations introduced by Zheng *et al.* (2019).

The GALMA is implemented by the following steps:

Step 1. The GA randomly generates the initial generation of parameter estimates. The selection, evaluation, crossover and mutation processes are repeated until the termination criterion is satisfied in which case the algorithm proceeds to Step 3. Otherwise upon reaching the maximum number of generations, the initial parameters in the LM algorithm use the best-fitting solution calculated with the GA.

Step 2. The trial solution and the SEE are updated along the steepest decent direction dp estimated with the LM algorithm. The detailed derivation of dp can be found in Zheng *et al.* (2019).

Step 3. The optimal solution is found when the calculated and measured precipitation satisfy a user-specified convergence criterion.

3 Results

3.1 Analysis based on univariate predictor variable

3.1.1 R/S analysis

Six test areas were selected among 27 meteorological stations as test examples. The R/S characteristic parameters of the monthly precipitation time series in the period 1988–2017 are listed in Table 1. Table 1 shows $H \in [0.785, 0.871]$, $0 \le F(z) < 1$, which demonstrates there is a strong long-term dependence, i.e., both that a high value in the precipitation series will probably be followed by another high value and that the precipitation values for some time into the future will also tend to be high. H is near 1, indicating the long-range dependence is relatively strong. In summary, the future monthly precipitation in Hubei Province are likely to be continuations of past trends.

Table 1. Rescaled range analysis parameters of monthly precipitation. R^2: coefficient of determination; H: Hurst index, $F(z)$: autocorrelation coefficient.

Test area	R^2	H	$F(z)$
1	0.729	0.785	0.485
2	0.716	0.795	0.505
3	0.701	0.815	0.548
4	0.712	0.793	0.501
7	0.678	0.871	0.672
8	0.814	0.856	0.638

3.1.2 LSTM model

The training and testing data of the LSTM model were from 1988–2012. The forecasting accuracy was determined by precipitation data from 2013–2017. The LSTM model was trained relying on monthly precipitation at the meteorological stations (test areas 1, 2, 3, 4, 7 and 8) as the datasets. There are several parameters that influence the LSTM neural networks, including the hidden units, learning rates and epochs. Computations were carried out with several parameter sets for a single-layer LSTM model to obtain the best performance parameters. Based on the narrow ranges of the RMSE and R^2, 200 was chosen as the number of hidden units, 600 for the epochs and 0.01 for the initial learning rate. The model was analysed during the training and testing phases to assess overfitting. During simulation, the loss functions for training and testing decreased and converged near zero, which suggests the model is a good fit. Finally, the accuracy of the prediction results increased by updating the network state with the observed values instead of the predicted values. The performances of the LSTM model for different test areas are listed in Table 2.

Table 2 shows the LSTM model performed well for the six test areas. The range of R^2 is 0.53–0.72 and the RMSE ranges from 54.17 to 85.28. The best predicted result is for Test Area 4 according to the highest R^2 value of 0.72. The smallest value of R^2 corresponds to Test Area 7 (0.53). The results reveal that LSTM model has good efficiency in forecasting time series of monthly precipitation in Hubei Province. Hence, the LSTM approach efficiently captures the characteristics of precipitation, which can also be visualized in the results between the observed and simulated data, as shown in Fig. 5 (taking Test Area 4 as an example).

3.1.3 SARIMA model

The dataset from 1988–2012 was applied in the estimation of model parameters. The dataset from 2013–2017 was used to determine the accuracy of the forecast, where the R^2, RMSE, MSE and MAE equal 0.68, 133.42, 17 800 and 86.96, respectively. Figure 6 displays the observed monthly precipitation data employed in this study for Test Area 4. Comparing the forecast results with the LSTM algorithm results demonstrates the LSTM algorithm is more suitable than SARIMA for forecasting monthly precipitation, having a lower RMSE and higher R^2.

The SARIMA(p,d,q) $(P,D,Q)[s]$ model in this study features a seasonal period of monthly precipitation equal to $S = 12$. Twelve-order seasonal differential processing was performed on the precipitation data to eliminate the seasonal period for the precipitation series. This is followed by first order difference was performed on the monthly precipitation data to make it a stationary sequence where $d = 1$, $D = 1$. Appropriate model parameters were estimated based on the PACF and ACF. The seasonal ACF and PACF are described in Fig. 7. Testing the different combinations of $(p,\ q)$ and $(P,\ Q)$ and further determining the best model with the lowest AIC and BIC yielded the best model parameters $p = 1$, $q = 3$ and $P = 1$, $Q = 2$. The optimal parameters of the SARIMA model are $(1,1,3)\ (1,1,2)_{12}$.

A diagnostic check was performed on the identified models (Fig. 8). The Box-Ljung testing has $P > 0.05$, which indicates the residual series of the model is white noise and shows that the model is reasonable. The model fit result is displayed in Fig. 8(c) with the observed data (blue line) and output data (red line). The 5-year precipitation forecast (2013–2017) is shown in Fig. 8(d).

Table 2. Model parameter estimates for six test areas. R^2: coefficient of determination; RMSE: root mean square error; MSE: mean square error; MAE: mean absolute error.

Test area	R^2	RMSE	MSE	MAE
1	0.68	78.08	6096.72	52.92
2	0.62	72.32	5230.48	55.58
3	0.64	75.54	5705.77	56.88
4	0.72	85.28	7271.95	59.35
7	0.53	54.17	2934.80	38.69
8	0.58	70.20	4928.51	51.31

3.2 Analysis based on multivariate predictor variables

3.2.1 BP-ANN and other machine learning algorithms

Machine learning techniques were applied to predict the monthly precipitation based on eight factors (LTD, altitude, NDVI, humidity, latitude, longitude, slope and month date) in the period 2005–2007. The training and testing data of the ML models were the first 33 months (see matrix ***B***). The accuracy was determined by the data of the last three months. Each method was cross-validated 10-fold. Several methods were compared, such as the ANNs and model trees (Table 3). The R^2 value varied in the range of 0.41 to 0.64, while the RMSE value was in the range 43.32 to 61.31. The model performance of the BP-ANN gives better R^2 (0.64) and RMSE (43.32) values when predicting monthly precipitation than other models. The next best model is the GPR model, which employs the rational quadratic kernel (rational quadratic GPR), with R^2 and RMSE values equal to 0.59 and 51.22, respectively. The poorest model is the linear SVR, with the lowest R^2 and highest RMSE. This comparison clearly indicates that monthly precipitations are effectively estimated by the BP-ANN model.

Table 3. Estimates of the model evaluation parameters for different algorithms.

Model	R^2	RMSE	MSE	MAE	Model	R^2	RMSE	MSE	MAE
BP-ANN	0.64	43.32	1876.5	31.90	Stepwise Linear	0.52	55.33	3061.2	37.29
Rational quadratic GPR	0.59	51.22	2623.8	33.13	Linear interactions	0.52	55.33	3061.4	37.21
Matern 5/2 GPR	0.59	51.23	2624.4	33.17	Medium tree	0.51	55.92	3127	36.61
Cubic SVR	0.59	51.28	2629.4	33.13	Coarse tree	0.47	57.84	3344.9	38.95
Boosted trees	0.58	51.42	2644.0	33.40	Linear regression	0.45	58.86	3464.4	41.31
Squared exponential GPR	0.58	51.43	2644.8	33.58	Coarse Gaussian SVR	0.45	59.31	3517.5	37.20
Exponential GPR	0.58	51.45	2647.0	33.07	Fine tree	0.43	60.01	3600.8	39.52
Bagged trees	0.58	51.82	2684.9	34.37	Fine Gaussian SVR	0.42	60.82	3699.5	40.75
Medium Gaussian SVR	0.58	51.84	2687.1	32.93	Robust linear	0.41	61.23	3749.6	39.67
Quadratic SVR	0.54	54.18	2935.6	34.87	Linear SVR	0.41	61.31	3759.3	39.63

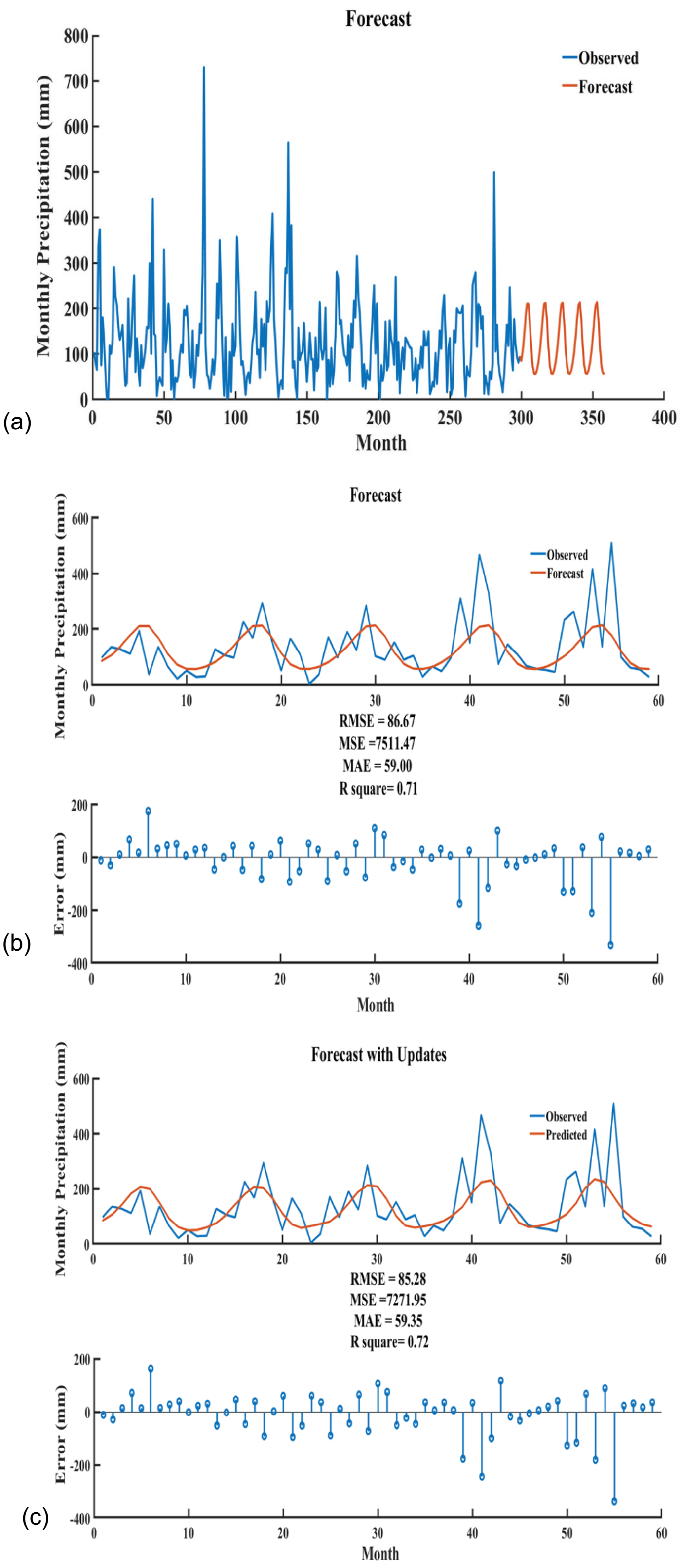

Figure 5. Results of the LSTM model for the monthly precipitation in Test Area 4, 1988–2017: (a) monthly precipitation prediction results of the LSTM model; (b) comparison between the predicted and observation data; and (c) updated network results comparing the prediction and observation data.

3.2.2 Simplified model for predicting monthly precipitation

3.2.2.1 Selecting model influence factors based on NMF.

Environmental and topographic factors have multiple interactions and influence precipitation. The correlation between these factors introduces duplication of information (multicollinearity). Matrix decomposition is a mapping of the original matrix to a subspace; the mapping is performed by approximating the low rank of the original matrix. The aim of NMF is to explore the relations between factors and precipitation. Altitude, slope, month date, latitude, longitude, LTD, NDVI, humidity and

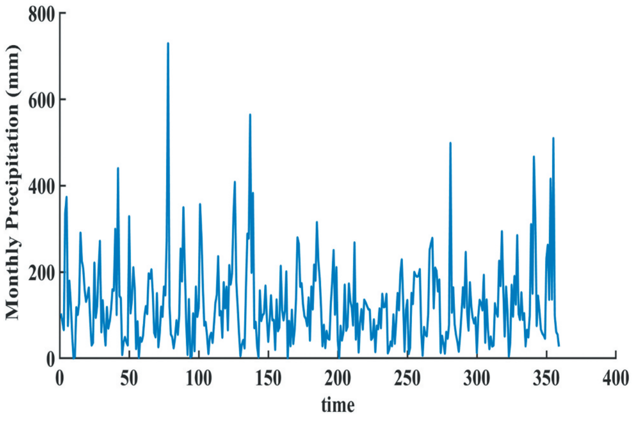

Figure 6. Observed monthly rainfall of raingauge station in Test Area 4.

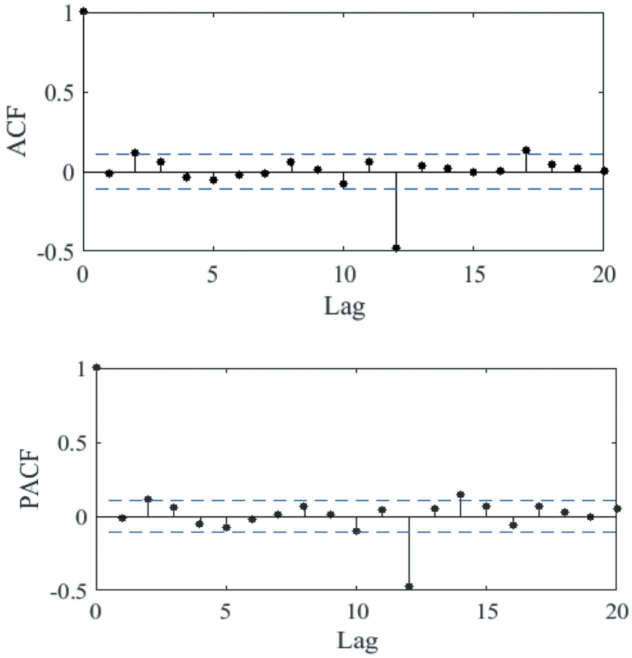

Figure 7. Seasonal ACF and PACF of monthly precipitation.

monthly precipitation data from 2005 to 2007 at 27 meteorological stations were normalized. The monthly date was transformed to make it consistent with the seasonal precipitation pattern of Hubei Province, i.e., May, June, July, August and September have larger weights, while the remaining months have smaller weights and the monthly date was normalized. A 927 rows × 9 columns matrix was constructed:

$$A = \begin{bmatrix} \mathrm{LTD}_1 & \cdots & \mathrm{longitude}_1 \\ \vdots & \ddots & \vdots \\ \mathrm{LTD}_n & \cdots & \mathrm{longitude}_n \end{bmatrix}$$

where the columns represent LTD, slope, rainfall, month date, NDVI, humidity, altitude, latitude and longitude. The rows are the corresponding factors arranged according to months and the meteorological stations from 1 to 27. Taking the mean value of the NMF 1000 times, the relations between the factors were expressed by a heat map (Fig. 9). The last two rows of Fig. 9(a) show that the influences of latitude and longitude on precipitation are negligible. Thus, the latitude and longitude factors were excluded from the original matrix and NFM was performed on the remaining factors of the matrix. The first row of Fig. 9(b) shows the LTD is closely related to the monthly date. The slope denotes a high correlation with altitude in the fourth row. The last row demonstrates there is a strong correlation of precipitation with LTD, humidity and altitude. The slope and month date were determined as redundant factors and they were excluded from the matrix. Only the internal relations between the selected factors were analysed; the selected factors are, the LTD, altitude, NDVI and humidity (see Fig. 9(c)). Humidity has a weak correlation with the NDVI and LTD (see second row). There is a high correlation between the LTD and NDVI in the first row of Fig. 9(c). The last row demonstrates the NDVI also correlates positively with altitude and humidity.

3.2.2.2 Influences of the selected parameters on monthly precipitation.

The empirical model was established according to the factors selected by NMF for predicting monthly precipitation in Hubei Province. The relations between a single factor and rainfall were analysed separately. Figure 10 displays a scatter gram of temperature and precipitation, which conforms to a Gaussian distribution. This is consistent with the conditions in Hubei Province. The rainfall in Hubei Province is concentrated in summer when temperature is higher.

The maximum, minimum and average values of each factor were adjusted; the thresholds of the other factors were established and the relations between a single factor and precipitation were compared. For example, when LTD < 10°C, altitude < 800 m and NDVI < 0.662, there is an exponential relation between relative humidity and precipitation (see Fig. 11). Humidity and precipitation follow the same trend when the threshold is changed. Exploring the association between altitude and precipitation indicates precipitation exhibits a decreasing trend with increasing altitude. The NDVI exhibits a linear relation with precipitation, while the slope of the trend line is flat. The evidence suggests altitude and NDVI have linear patterns of association with precipitation.

3.2.2.3 Semi-empirical model.

The following formula was fitted based on the analysis of relations between factors:

$$Y = f(T; P_1, P_2) \cdot P_3(\mathrm{Alt} + P_4) \cdot (\mathrm{NDV} + P_5) \cdot e^{H \cdot P_6} \qquad (13)$$

where Y denotes the monthly precipitation and f represents the Gaussian membership function $f(T; P_1, P_2) = \exp((-(T - P_1)^2)/2P_2{}^2)$. T represents the LTD (°C) and P_1 and P_2 denote the mean and the standard deviation of the Gaussian distribution, respectively. The terms Alt, H and NDV denote the altitude (m), humidity and NDVI, respectively, and P_3, P_4, P_5 *and* P_6 are constants.

Parameter estimation based on the nonlinear model (13) was carried out with a hybrid algorithm named GALMA (Zheng *et al.* 2019) (Section 2.3.3.2). The objective function for parameter estimation is the MSE between the observed data and predicted precipitation obtained with Equation (13). Figure 12 presents the convergence history during the estimation of parameters. The GA is applied first and the LM starts its search with the 20th generation of the GA. It is seen in Fig. 12 the improvement achieved by

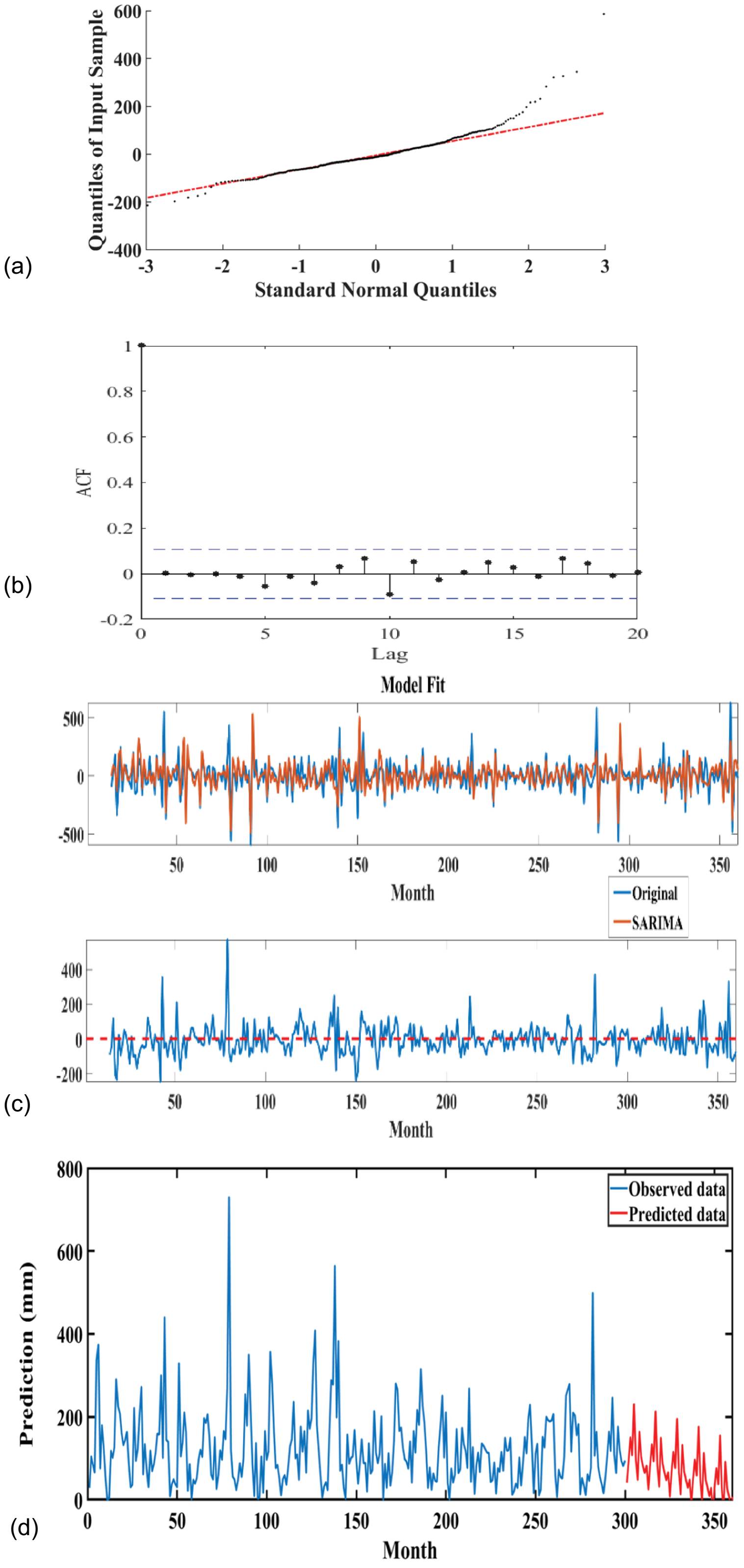

Figure 8. SARIMA model diagnostic check and model fit results: (a) residual line chart of the ARIMA model; (b) ACF of the residuals; (c) model fit and residual plot; and (d) predicted precipitation.

the GALMA. After 120 iterations of the LM, the value of the objective function converges with a set of near-optimal parameters.

The optimal parameters obtained by the GALMA are $P_1 = 26.558$, $P_2 = 43.512$, $P_3 = 1.442 \times 10^{-4}$, $P_4 = 1.335 \times 10^{4}$, $P_5 = 0.120$, and $P_6 = 6.203$. The value of R^2 is 0.65 (see Figure 13), RMSE is 41.81, MSE is 1747.76 and MAE is 29.11. Thus, the predictive skill of the proposed semi-empirical model is at least as good as that of machine learning algorithms.

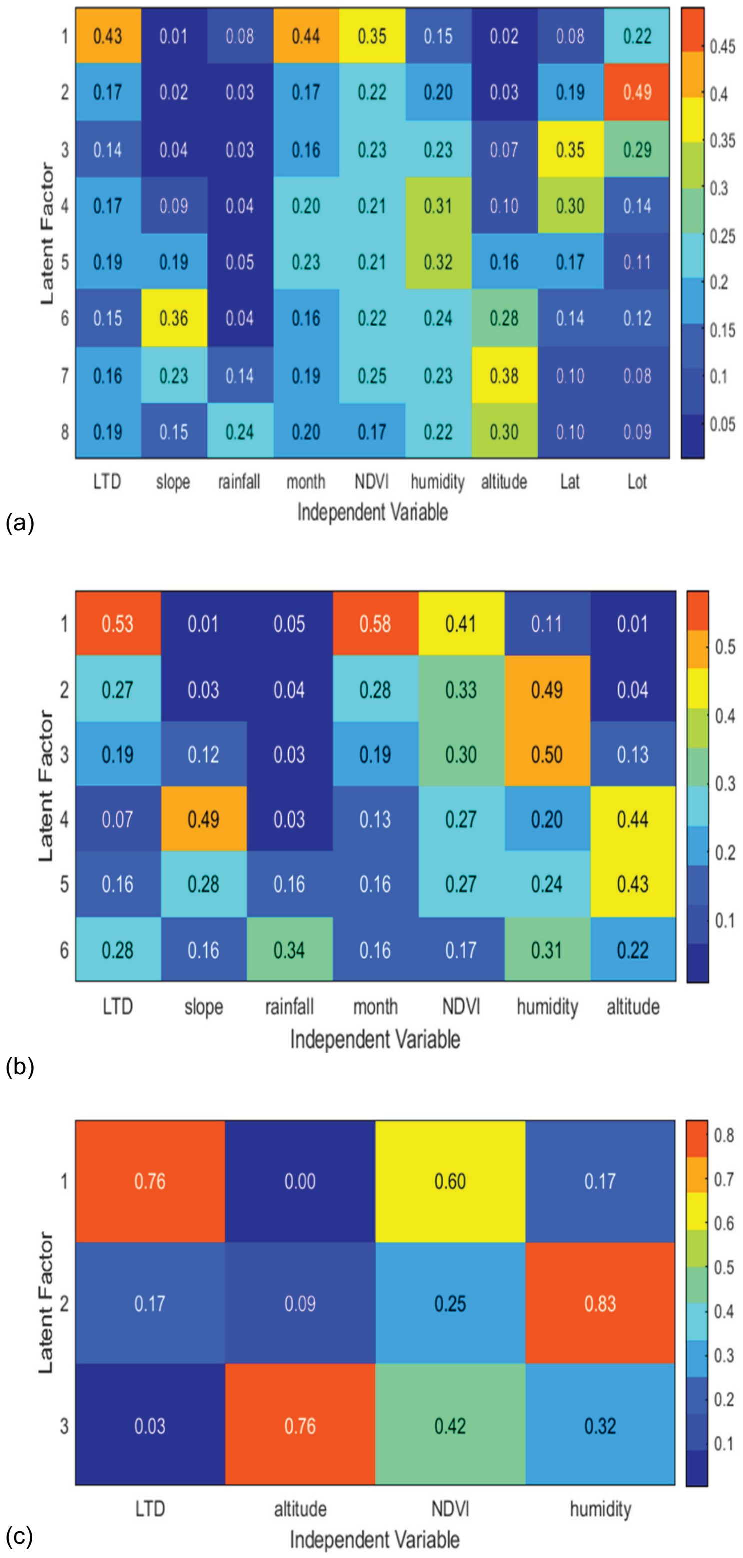

Figure 9. NMF heat maps of: (a) nine NMF factors; (b) seven NMF factors; and (c) four NMF factors.

4 Discussion and conclusions

This paper relied on monthly rainfall time series featuring strong long-term dependence in the period 1988–2017. Results indicated that there was a persistent characteristic of precipitation in Hubei Province. This paper also compared the prediction results obtained with the one-step ahead LSTM and SARIMA algorithms based on monthly time series. The prediction accuracy of the LSTM model was better than that of the SARIMA model. It may depend on the sample sizes (Papacharalampous *et al*. 2018a, 2018b).

Environmental and topographic factors were introduced in the prediction models. The performances of the machine learning

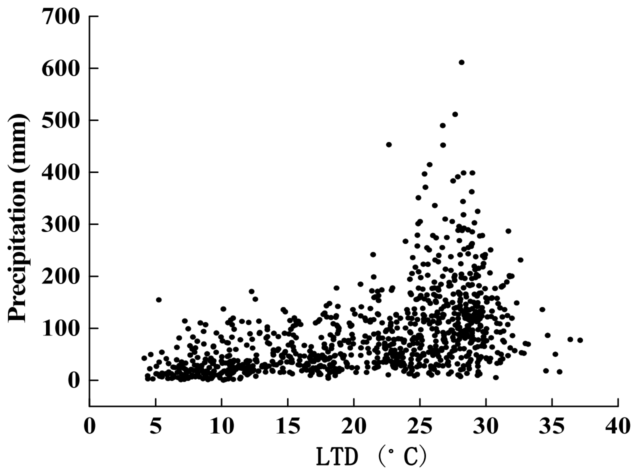

Figure 10. Relationship between the LTD and monthly precipitation.

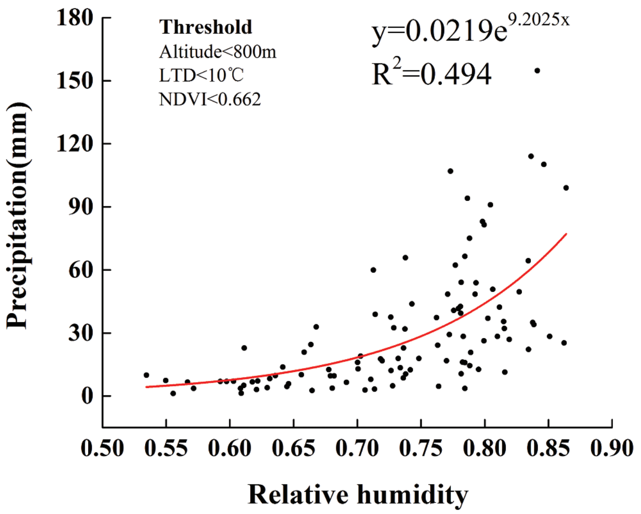

Figure 11. Fitted function between relative humidity and monthly precipitation.

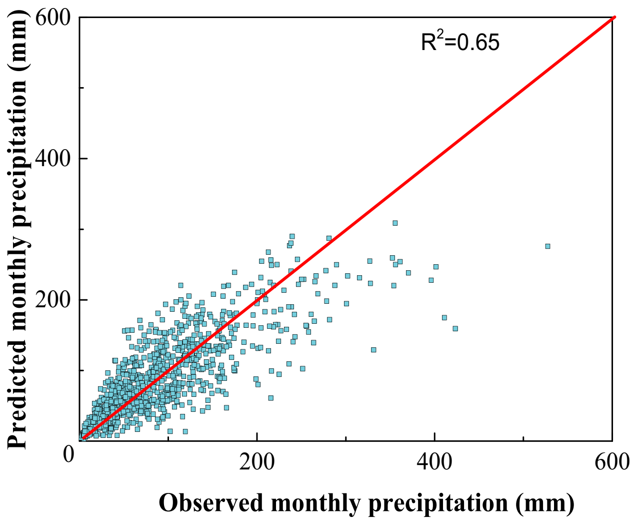

Figure 13. Prediction results of the semi-empirical model.

algorithms for monthly precipitation were assessed based on the environmental factors. The results implied the BP-ANN has the best performance among the algorithms.

A novel approach based on developing the semi-empirical model was proposed in this study for the prediction of monthly precipitation. The approach included combining NMF, numerical analysis and the GALMA parameter optimization algorithm. NMF was applied to explore the potential connection between the influence factors and monthly precipitation. The results revealed that relative humidity, daily surface temperature, altitude and NDVI were the main governing factors for predicting monthly precipitation and they were chosen in the model development for this paper. Determining the optimal parameters of the prediction model by GALMA combines the advantages of the GA and LM algorithms.

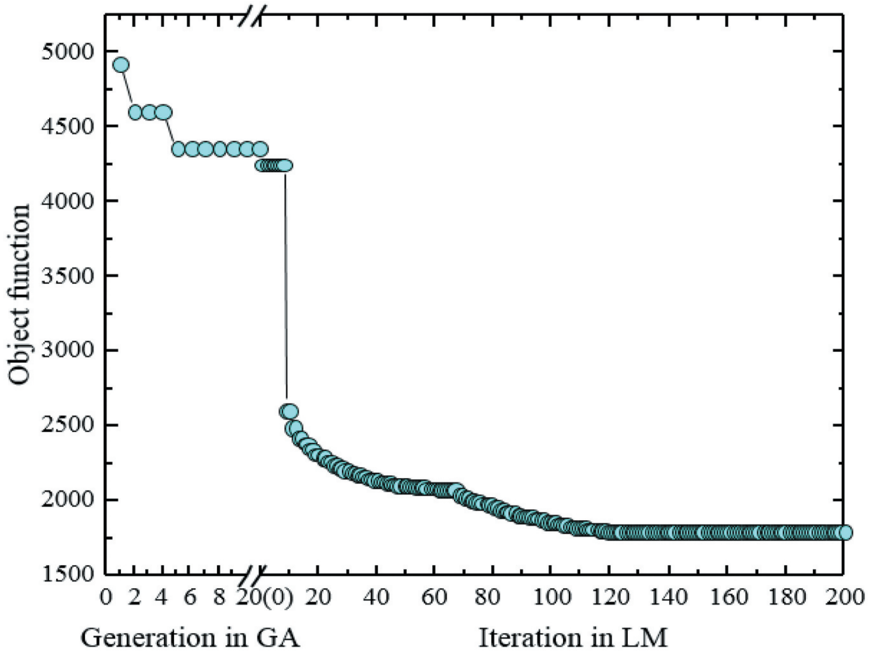

Figure 12. Convergence rate of the GALMA for estimation.

Future work will evaluate the suitability of the semi-empirical model for other areas besides Hubei Province. In addition, more environmental factors could be entertained in the prediction model to select the most relevant factors and parameters. Future research will focus on improving the accuracy of the prediction model and comparing the performance of enhanced machine learning algorithms for predicting monthly rainfall.

Acknowledgements

We are grateful to the editorial team and three reviewers of professor Andreas Langousis, Dr Hristos Tyralis and another anonymous reviewer for their efforts to improve this manuscript. This work was partly supported by the China Scholarship Council (NO.201806270111).

Disclosure statement

The authors report no potential conflicts of interest.

Funding

This study was supported by the National Key Research and Development Program of China [2016YFC0803106].

ORCID

Huihui Zhang http://orcid.org/0000-0003-4920-8868
Qingyun Du http://orcid.org/0000-0003-4615-2029

References

Abbot, J. and Marohasy, J., 2014. Input selection and optimization for monthly rainfall forecasting in Queensland, Australia, using artificial neural networks. *Atmospheric Research*, 138, 166–178. doi:10.1016/j.atmosres.2013.11.002

Alireza, H. and Peter, S., 2018. *Application of soft computing and intelligent methods in geophysics*. 1st ed, 25–26. New York: Springer International Publishing. doi: 10.1007/978-3-319-66532-0

Bahram, C., *et al.*, 2018. Precipitation forecasting using classification and regression trees (CART) model: a comparative study of different approaches. *Environmental Earth Sciences*, 77, 314. doi:10.1007/s12665-018-7498-z

Bao, Y.J., *et al.*, 2007. Study on the spatial differences and its time lag effect on climatic factors of the vegetation in the longitudinal range-gorge region. *Chinese Science Bulletin*, 52 (S2), 42–49. doi:10.1007/s11434-007-7005-5

Beran, J., *et al.*, 2013. *Long-memory processes*. New York: Springer-Verlag Berlin Heidelberg. doi:10.1007/978-3-642-35512-7

Box, G.E.P., *et al.*, 2015. *Time series analysis: forecasting and control*. 5th ed. Hoboken, New Jersey: John Wiley and Sons Inc.

Breiman, L., *et al.*, 1984. *Classification and regression trees*. New York: Chapman & Hall.

Breiman, L., 1996. Bagging predictors. *Machine Learning*, 24 (2), 123–140. doi:10.1007/BF00058655

Brunet, J.P., *et al.*, 2004. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 101 (12), 4164–4169. doi:10.1073/pnas.0308531101

Cao, X.M., *et al.*, 2011. Response of vegetation to temperature and precipitation in Xinjiang during the period of 1998–2009. *Journal of Arid Land*, 3 (2), 94–103. doi:10.3724/SP.J.1227.2011.00094

Catalogna, M., *et al.*, 2012. Artificial neural networks based controller for glucose monitoring during clamp test. *PloS One*, 7 (8), e44587. doi:10.1371/journal.pone.0044587

Chaparro, L.F., 2019. *Signals and systems using MATLAB*. 3rd ed. London, UK: Academic Press.

Chen, J.D. and Boccelli, D.L., 2018. Real-time forecasting and visualization toolkit for multi-seasonal time series. *Environmental Modelling and Software*, 105, 244–256. doi:10.1016/j.envsoft.2018.03.034

Chinchorkar, S.S., Patel, G.R., and Sayyad, F.G., 2012. Development of Monsoon Model for Long Range Forecast Rainfall Explored for Anand (Gujarat-India). *International Journal of Water Resources and Environmental Engineering*, 4 (11), 322–326. doi:10.5897/IJWREE11.097

Colette, A., Vautard, R., and Vrac, M., 2012. Regional climate downscaling with prior statistical correction of the global climate forcing. *Geophysical Research Letters*, 39 (13). doi:10.1029/2012GL052258

Dirks, K.N., *et al.*, 2003. A semi-empirical model for predicting the effect of changes in traffic flow patterns on carbon monoxide concentrations. *Atmospheric Environment*, 37 (19), 2719–2724. doi:10.1016/S1352-2310(03)00156-0

Eni, D. and Adeyeye, F.J., 2015. Seasonal ARIMA modelling and forecasting of rainfall in Warri town, Nigeria. *Journal of Geoscience and Environment Protection*, 03 (6), 91–98. doi:10.4236/gep.2015.36015

Fan, R.E., *et al.*, 2005. Working set selection using second order information for training support vector machines. *The Journal of Machine Learning Research*, 6, 1871–1918.

Fan, R.E., *et al.*, 2006. A study on SMO-type decomposition methods for support vector machines. *IEEE Transactions on Neural Networks*, 17, 893–908. doi:10.1109/TNN.2006.875973

Hashemifard, S.A., *et al.*, 2019. Predicting the rarefied gas flow through circular nano/micro short tubes: a semi-empirical model. *Vacuum*, 164, 18–28. doi:10.1016/j.vacuum.2019.02.044

Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. *Neural Computation*, 9 (8), 1735–1780. doi:10.1162/neco.1997.9.8.1735

Jedrzejek, B., *et al.*, 2013. Vegetation pattern of mountains in west Greenland - a baseline for long-term surveillance of global warming impacts. *Plant Ecology & Diversity*, 6 (3–4), 405–422. doi:10.1080/17550874.2013.802049

Jha, M.K., *et al.*, 2006. Evaluation of traditional and non-traditional optimization techniques for determining well parameters from step-drawdown test data. *Journal of Hydrologic Engineering*, 11 (6), 617–630. doi:10.1061/(ASCE)1084-0699(2006)11:6(617)

Karmakar, S., Senjuti, G., and Surajit, C., 2019. Exploring the pre- and summer-monsoon surface air temperature over eastern India using Shannon entropy and temporal Hurst exponents through rescaled range analysis. *Atmospheric Research*, 217, 57–62. doi:10.1016/j.atmosres.2018.10.007

Khan, N., *et al.*, 2019. Prediction of heat waves in Pakistan using quantile regression forests. *Atmospheric Research*, 221, 1–11. doi:10.1016/j.atmosres.2019.01.024

Kumar, D., *et al.*, 2019. Forecasting monthly precipitation using sequential modelling. *Hydrological Sciences Journal*, 64 (6), 690–700. doi:10.1080/02626667.2019.1595624

Kumar, S.V. and Lelitha, V., 2015. Short-term traffic flow prediction using seasonal ARIMA model with limited input data. *European Transport Research Review*, 7 (3), 21. doi:10.1007/s12544-015-0170-8

LeCun, Y., Bengio, Y., and Hinton, G., 2015. Deep learning. *Nature*, 521 (7553), 436–444. doi:10.1038/nature14539

Lee, D.D. and Seung, H.S., 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401 (6755), 788–791. doi:10.1038/44565

Liu, Q.J., *et al.*, 2013. Modeling the daily suspended sediment concentration in a hyperconcentrated river on the Loess Plateau, China, using the Wavelet-ANN approach. *Geomorphology*, 186, 181–190. doi:10.1016/j.geomorph.2013.01.012

Ljung, G.M. and Box, G.E.P., 1978. On a measure of lack of fit in time series models. *Biometrika*, 65, 297–303. doi:10.1093/biomet/65.2.297

Loh, W.Y., 2002. Regression Trees with unbiased variable selection and interaction detection. *Statistica Sinica*, 12, 361–386.

López-Lambraño, A.A., *et al.*, 2018. Spatial and temporal Hurst exponent variability of rainfall series based on the climatological distribution in a semiarid region in Mexico. *Atmósfera*, 31 (3), 199–219. doi:10.20937/ATM.2018.31.03.02

Madsen, K., Nielsen, H., and Tingleff, O. 2004. *Methods for nonlinear least squares problems*. Technical Report, Informatics and Mathematical Modelling, Technical University of Denmark, Kongens Lyngby, Denmark.

Mandelbrot, B. and Wallis, J.R., 1969. Robustness of the rescaled range R/S in the measurement of noncyclic long-run statistical dependence. *Water Resources Research*, 5, 967–988. doi:10.1029/WR005i005p00967

Martin, M., *et al.* 2017. Data mining in cloud usage data with MATLAB's statistics and machine learning toolbox. *2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI), Herl'any*.

Mehdizadeh, S., Behmanesh, J., and Khalili, K., 2018. New approaches for estimation of monthly rainfall based on GEP-ARCH and ANN-ARCH hybrid models. *Water Resources Management*, 32 (2), 527–545. doi:10.1007/s11269-017-1825-0

Mekanik, F., *et al.*, 2013. Multiple regression and artificial neural network for long-term rainfall forecasting using large scale climate modes. *Journal of Hydrology*, 503, 11–21. doi:10.1016/j.jhydrol.2013.08.035

Milanic, S., *et al.*, 1998. A semi-empirical and a network model of a batch plant dynamics for flexible recipes control. *IFAC Proceedings Volumes*, 31 (11), 289–294. doi:10.1016/S1474-6670(17)44943-3

Murthy, K.V.N., Saravana, R., and Kumar., K.V., 2018. Modeling and forecasting rainfall patterns of southwest monsoons in north–east India as a SARIMA process. *Meteorology and Atmospheric Physics*, 130 (1), 99–106. doi:10.1007/s00703-017-0504-2

Nourani, V., Kisi, Ö., and Komasi, M., 2011. Two hybrid artificial intelligence approaches for modeling rainfall-runoff process. *Journal of Hydrology*, 402 (1–2), 41–59. doi:10.1016/j.jhydrol.2011.03.002

Novak, M. and Mammone, R., 2001. Improvement of non-negative matrix factorization based language model using exponential models. *IEEE Workshop on Automatic Speech Recognition and Understanding*. Italy: ASRU '01, 190–193. doi:10.1109/ASRU.2001.1034619

Paatero, P. and Tapper, U., 1994. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5 (2), 111–126. doi:10.1002/env.3170050203

Pal, L., *et al.*, 2019. Regional scale analysis of trends in rainfall using nonparametric methods and wavelet transforms over a semi-arid region in India. *International Journal of Climatology*, 39 (5), 2737–2764. doi:10.1002/joc.5985

Papacharalampous, G., *et al.*, 2018a. Predictability of monthly temperature and precipitation using automatic time series forecasting methods. *Acta Geophysica*, 66 (4), 807–831. doi:10.1007/s11600-018-0120-7

Papacharalampous, G., *et al.*, 2018b. One-step ahead forecasting of geophysical processes within a purely statistical framework. *Geoscience Letters*, 5 (1), 12. doi:10.1186/s40562-018-0111-1

Papacharalampous, G., *et al.*, 2019. Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes. *Stochastic Environmental Research and Risk Assessment*, 33 (2), 481–514. doi:10.1007/s00477-018-1638-6

Pedregal, D.J., 2019. Time series analysis and forecasting with ECOTOOL. *PloS One*, 14 (10), e0221238. doi:10.1371/journal.pone.0221238

Ramana, R.V., *et al.*, 2013. Monthly rainfall prediction using wavelet neural network analysis. *Water Resources Management*, 27 (10), 3697–3711. doi:10.1007/s11269-013-0374-4

Rasmussen, C.E. and Williams, C.K.I., 2006. *Gaussian processes for machine learning*. Cambridge, Massachusetts: The MIT Press.

Reed, P., Minsker, B., and Goldberg, D.E., 2000. Designing a competent simple genetic algorithm for search and optimization. *Water Resources Research*, 36 (12), 3757–3761. doi:10.1029/2000WR900231

Samuel, M.P. and Jha, M.K., 2003. Estimation of aquifer parameters from pumping test data by genetic algorithm optimization technique. *Journal of Irrigation and Drainage Engineering*, 129 (5), 348–359. doi:10.1061/(ASCE)0733-9437(2003)129:5(348)

Sapankevych, N.I. and Sankar, R., 2009. Time series prediction using support vector machines: a survey. *IEEE Computational Intelligence Magazine*, 4 (2), 24–38. doi:10.1109/MCI.2009.932254

Shi, X.J., *et al.*, 2015. Convolutional LSTM network: a machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems*, 28, 802–810.

Shumway, R.H. and Stoffer, D.S., 2017. *Time series analysis and its applications with R examples 4th*. New York: Springer.

Voyant, C., *et al.*, 2017. Machine learning methods for solar radiation forecasting: a review. *Renewable Energy*, 105, 569–582. doi:10.1016/j.renene.2016.12.095

Warmuth, M., *et al.* 2006. Totally corrective boosting algorithms that maximize the margin. *Conference: Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA*.

Wu, Q.H., *et al.*, 2009. Effects of topography and urban heat circulation on a meso-βtorrential rain in Beijing area. *Meteorology Monthly*, 35, 58–64. (in Chinese).

Zhang, J.Y., *et al.*, 2003. The influence of vegetation cover on summer precipitation in China: a statistical analysis of NDVI and climate data. *Advances in Atmospheric Sciences*, 20 (6), 1002–1006. doi:10.1007/BF02915523

Zheng, G., *et al.*, 2019. Estimation of the hydraulic parameters of leaky aquifers based on pumping tests and coupled simulation/optimization: verification using a layered aquifer in Tianjin, China. *Hydrogeology Journal*, 27, 3081–3095. doi:10.1007/s10040-019-02021-z