Journal of Environmental Management 222 (2018) 190–206

Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: www.elsevier.com/locate/jenvman

Research article

Novel hybrid linear stochastic with non-linear extreme learning machine methods for forecasting monthly rainfall a tropical climate

Mohammad Zeynoddin[a], Hossein Bonakdari[a,b,*], Arash Azari[c], Isa Ebtehaj[a,b], Bahram Gharabaghi[d], Hossein Riahi Madavar[e]

[a] Department of Civil Engineering, Razi University, Kermanshah, Iran
[b] Environmental Research Center, Razi University, Kermanshah, Iran
[c] Department of Water Engineering, Razi University, Kermanshah, Iran
[d] School of Engineering, University of Guelph, Guelph, Ontario, NIG 2W1, Canada
[e] Department of Water Engineering, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran

ARTICLE INFO

Keywords:
Forecast
Hybrid model
Rainfall
Spectral analysis
Tropical climate
Uncertainty

ABSTRACT

A novel hybrid approach is presented that can more accurately predict monthly rainfall in a tropical climate by integrating a linear stochastic model with a powerful non-linear extreme learning machine method. This new hybrid method was then evaluated by considering four general scenarios. In the first scenario, the modeling process is initiated without preprocessing input data as a base case. While in other three scenarios, the one-step and two-step procedures are utilized to make the model predictions more precise. The mentioned scenarios are based on a combination of stationarization techniques (i.e., differencing, seasonal and non-seasonal standardization and spectral analysis), and normality transforms (i.e., Box-Cox, John and Draper, Yeo and Johnson, Johnson, Box-Cox-Mod, log, log standard, and Manly). In scenario 2, which is a one-step scenario, the stationarization methods are employed as preprocessing approaches. In scenario 3 and 4, different combinations of normality transform, and stationarization methods are considered as preprocessing techniques. In total, 61 sub-scenarios are evaluated resulting 11013 models (10785 linear methods, 4 nonlinear models, and 224 hybrid models are evaluated). The uncertainty of the linear, nonlinear and hybrid models are examined by Monte Carlo technique. The best preprocessing technique is the utilization of Johnson normality transform and seasonal standardization (respectively) (R^2 = 0.99; *RMSE* = 0.6; *MAE* = 0.38; *RMSRE* = 0.1, *MARE* = 0.06, *UI* = 0.03 & *UII* = 0.05). The results of uncertainty analysis indicated the good performance of proposed technique (d-factor = 0.27; 95PPU = 83.57). Moreover, the results of the proposed methodology in this study were compared with an evolutionary hybrid of adaptive neuro fuzzy inference system (ANFIS) with firefly algorithm (ANFIS-FFA) demonstrating that the new hybrid methods outperformed ANFIS-FFA method.

1. Introduction

Rainfall has a prominent impact on numerous applications in water resource planning and management, particularly in the tropical climate regions of the globe (Luk et al., 2001; Gazendam et al., 2016; Ng et al., 2017). Due to large variations in rainfall events, especially in tropical areas, accurate rainfall forecasting is crucial in water resource management such as potable water demand, ground water availability, surface water hydrology, hydropower, irrigation water demand and flood control. However, due to the absence of accessible information on the climatic, topographic, physiographic, soils, and land use datasets, it is not straightforward to predict rainfall using a physical-based model.

Some researcher have successfully employed the stochastic models based on Box and Jenkins (1976) methodology (i.e. auto-regressive moving average, ARMA, auto-regressive integrated moving average, ARIMA, and seasonal ARIMA, SARIMA, and etc) are the most popular models in hydrological time series forecasting (Kumar and Goyal, 2011; Jian et al., 2012; Valipour et al., 2013; Pektaş and Cigizoglu, 2013; Mirzavand and Ghazavi, 2015; Nieto et al., 2018) especially rainfall (Osarumwense, 2013; Modarres and Ouarda, 2013; Wang et al., 2015; Valipour, 2015). The SARIMA model could be employed just for time series with stationary at variance. A significant disadvantage of SARIMA is the linear character of the model so that it cannot recognize and model the extreme values of a time series accurately.

In more recent studies, to overcome the limitation of stochastic models for hydrological time series especially rainfall forecasting,

* Corresponding author. Department of Civil Engineering, Razi University, Kermanshah, Iran.
E-mail address: bonakdari@yahoo.com (H. Bonakdari).

https://doi.org/10.1016/j.jenvman.2018.05.072
Received 5 February 2018; Received in revised form 20 May 2018; Accepted 22 May 2018

numerous artificial intelligence (AI) techniques such as artificial neural network (ANN) (Zeroual et al., 2016; Babel et al., 2017; Shi et al., 2018), adaptive neuro-fuzzy inference system (ANFIS) (Awan and Bae, 2014; Yaseen et al., 2018), support vector machines (SVM) (Hong and Pai, 2007; Devak et al., 2015; Tehrany et al., 2015), Multivariate Adaptive Regression Spline (MARS) (Adamowski et al., 2012), Gene-Expression Programming (GEP) (Phukoetphim et al., 2016; Mehdizadeh et al., 2017) have been introduced.

The principal shortcoming of the ANN method has to do with deciding the optimal structure and appropriate size. Controlling training parameters for convergence and learning is remarkably troublesome errand. Moreover, the main drawback of this technique is the over-training which results in unstable forecasting abilities. The major disadvantages of ANFIS are the high computational cost and the complexity related to if-then rules and membership function optimization (Mazari and Rodriguez, 2016). In modeling a multi-input-single-output problem using SVM, the parameters of this method should be adjusted to gain the optimum results. Indeed, the value of these parameters has a significant impact on model accuracy. Another limitation of the SVM is the selection of proper kernel function so that there is not a uniform standard to find the best kernel function for a specific problem.

To overcome the limitation of AI-based techniques, several hybrid methods were introduced. The hybrid methods are categorized into two groups: the use of preprocessing techniques and optimization methods. In the field of optimization methods, different optimization algorithms such as evolutionary algorithms (EAs) were employed to optimize the AI-based method. For example, the EA is combined with ANN and ANFIS to optimize the weights and membership function values, respectively, to overcome the weakness of traditional algorithms such as backpropagation (BP) (Azimi et al., 2016; Ghorbani et al., 2017; Yaseen et al., 2018). Moreover, the EA optimized the adjustable variables of SVM techniques which were determined through a trial and error process (Ebtehaj and Bonakdari, 2016; Ebtehaj et al., 2018). In other hybrid methods, the preprocessing techniques are employed which transform the main time series to new series or decompose it using wavelet transforms or ensemble empirical mode decomposition to different time series by considering the time series specifications as details.

Recently, a few studies in the field of hydrological time series forecasting were implemented based on recognition hydrological time series components. Moeeni and Bonakdari (2017) express that dam reservoir inflow is an extreme seasonal time series, and it is not purely linear or nonlinear. Therefore, the use of a hybrid framework to consider the linear and nonlinear terms of a hydrological time series simultaneously could be an alternative approach to overcome the drawback of each linear and nonlinear techniques. They presented a hybrid technique based on a combination of SARIMA and ANN as linear and nonlinear techniques, respectively. Following later, Moeeni et al. (2017a) and Moeeni et al. (2017b) examined the performance of GEP and ANFIS as two favorite nonlinear techniques in the proposed methodology. They found that the proposed methodology could be considered as a robust technique in forecasting of dam reservoir inflow.

The primary goal of this study is to develop a novel more accurate hybrid linear-nonlinear framework for forecasting monthly rainfall in tropical climatic regions. In this method, four different stationarization techniques (i.e. differencing, seasonal and non-seasonal standardization and spectral analysis) (Salas et al., 1980; Moeeni et al., 2017c) and eight normalization methods, (i.e. Box-Cox and Box-Cox-Modified, John-Draper, Yeo-Johnson, Johnson, Log and Log standard, and Manly) (Box and Cox, 1964; John and Draper, 1980; Yeo and Johnson, 2000; Johnson, 1949; Salas et al., 1980; Manly, 1976) are considered in four different scenarios. In the first scenario, the hybrid model is employed without any preprocessing while in the others, a one-step preprocessing (scenario 2) and two two-steps preprocessing (scenario 3 and 4) are considered in hybrid modeling. In addition, the existence of rainfall time series components such as jump, trend and period are examined by statistical tests. Generally, 61 sub-scenarios are produced which resulted, 10785 linear, 4 nonlinear and 224 hybrid models. These models are scrutinized to find the optimum method. Moreover, the uncertainty of the proposed hybrid method, linear and nonlinear techniques are evaluated using Monte Carlo simulation techniques.

The remnant of this paper is structured as follows: The study area and the characteristics of rainfall time series is provided in section 2. Section 3 presents the theoretical framework. Following, sub-sections 3.1 through 3.2 offer a brief description of stochastic models and ELM technique, and sub-section 3.3 provides the step by step process employed to develop the integrated stochastic model with nonlinear techniques for monthly rainfall modeling. Section 4 provides the performance evaluation criteria. Section 5 then presents the results of the proposed hybrid linear-nonlinear methodology. Finally, section 6 includes the concluding remarks and opening for a future study in this field.

2. Study area and hydrological data

In this study, the monthly rainfall data for Pahang basin, Malaysia has been collected as a case study to validate the proposed new hybrid linear-nonlinear methodology. Pahang basin is the third largest state after Sabah and Sarawak states and the largest in peninsular Malaysia. It has an average area of around 36,137 km^2 (comprising 11% of the total land area of Malaysia). This watershed is surrounded by Johor, Kelantan, Negeri Sembilan, and Selangor, the South China Sea and Terengganu from the south, north, west, and east, respectively (Fig. 1). Pahang watershed has received high rainfall through northeast monsoon period with almost 40% of total monthly rainfall (JMM, 2010). Extreme rainfall activated by northeast monsoon is the principle factor in extreme flood events in this basin (DID, 2005; DID, 2009). The excessive river flood flows in Pahang basin has resulted in major modification of river channel size and damage to floodplain properties (Camporeale et al., 2007). This state plays a vital role in the region's economy, particularly in agriculture, manufacturing and service fields and supports the entire east coast region of peninsular Malaysia. Highly variable nature of rainfall events affect river floods every year in this area. Thus, it is of great importance to achieve more accurate rainfall prediction to reservoir management planning to decrease the risk of flood damages.

The annual rainfall in this watershed varies between 1609 mm and 2132 mm for Temerloh and Lubuk Paku stations (respectively). Intense precipitations in this basin occur from November to March. June and July are also the driest periods. In this research study, a total of 180 monthly rainfall records of Pahang watershed time series data between the periods from 2000 to 2014 have been employed to build predictive models. This data was measured at the most downstream station in the basin, Lobok Pakou Station. The data was recorded daily and averaged monthly for this study. Fig. 1 indicates the primary statistical indices of collected data within total, training and testing stages for Pahang basin. In this figure is also indicated the variation in rainfall time series for the period between January 2000 and December 2014. It is notable that the observed data from January 2000 to December 2011 (i.e., the 70% of total dataset or a full of 144 month records) were allocated for training stages and the data from January 2011 to December 2014 (i.e., the 30% of total dataset or a full of 36 month records) were assigned to testing stage. Except for minimum, maximum, which are extreme values observed in the hydrological events, skewness and kurtosis and the value of other statistical characterizes were found to be approximately similar for testing and training datasets.

3. Theoretical framework

In this section, a brief explanation about the stochastic model as a linear and extreme learning machine (ELM) as nonlinear models are presented. Afterward, the new hybrid linear-nonlinear method which is