# Daily river flow prediction based on Two-Phase Constructive Fuzzy Systems Modeling: A case of hydrological – meteorological measurements asymmetry

Bassam Bou-Fakhreddine, Sara Abou Chakra, Imad Mougharbel, Alain Faye, Yann Pollet

Research Paper

# Daily River Flow Prediction Coupled with Data Processing Techniques: A Comparative Study between Constructive Fuzzy Systems and Autoregressive Models

*Bassam BOU-FAKHREDDINE[a,e], Sara ABOU CHAKRA[b], Imad MOUGHARBEL[c], Alain FAYE[d], Yann POLLET[e]*

[a] *Doctoral School of Science and Technology, Lebanese University, Hadath, Lebanon*
[b] *University Institute of Technology, Lebanese University, Aabey, Lebanon*
[c] *Faculty of Engineering, Lebanese University, Hadath, Lebanon*
[d] *Ecole Nationale Superieure d'Informatique pour l'Industrie et l'Entreprise, Evry, France*
[e] *CEDRIC, Conservatoire National des Arts et Metiers, Paris, France*

**Abstract**

Daily river flow forecast is an essential step for real-time hydro-power reservoir operation. The purpose of the flow forecast is to assist in the decision-making process in order to ensure optimal and reliable operational policy. The paper presents, in a region where meteorological and hydrological data are insufficient, inaccessible and sometimes unreliable, a data-driven model based on Constructive Fuzzy Systems. The model is capable of exploiting the available data with high prediction efficiency was compared to an Autoregressive method. A case study was applied to Litani River in the Bekaa Valley - Lebanon using 4 years of rainfall, temperature, and river flow daily measurements. A reference Auto-Regressive (AR) model, a classical Constructive Fuzzy System Modeling (C-FSM) and the Constructive Fuzzy System Modeling coupled with Moving Average (C-FSM_MA) filter are trained. Upon testing, the last two models have shown primarily competitive performance and accuracy with the ability of preserving the day-to-day variability up to 12 days ahead. In fact, for the longest lead period, the models AR, C-FSM and C-FSM_MA were able of explaining respectively 75%, 79.5% and 84.3% of the actual river flow variation. These results indicate that Moving Average (MA) filter provides a supportive pre-processing tool in the process of streamflow forecasting.

*Keywords:* Forecasting daily river flow, Data pre-processing, Fuzzy systems, Autoregressive model, Litani River

## 1. Introduction

Accurate prediction of river flow is of vital importance for efficient reservoir water management and control. However, forecasting river flow remains one of the very difficult issues in hydrological sciences because it is characterized by a dynamic, uncertain and nonlinear problem (Huamani et al., 2011). This problem deals with a system that receives thousands of inputs interacting in a complex and noisy environment.

Over the past few decades, several types of stochastic models have been suggested for hydrological time series modeling such as Box and Jenkins (Box and Jenkins, 1970) methods for Auto-Regressive (AR), Auto-Regressive Moving Average (ARMA), Auto-Regressive Integrated Moving Average (ARIMA), Auto-Regressive Moving Average with Exogenous inputs (ARMAX) models. They were generally utilized in the linear sense for estimating future river flow. Later on, several studies were dedicated to the formularization and development of nonlinear river flow models that aim to improve the quality of hydrological forecasting. In fact, (Porporato and Ridolf, 2001) dealt with local linear models with time-dependent parameters, whereas (Dibike and Solomatine, 2001) and (Pulido-Calvo and Portela, 2007) have considered data-driven nonlinear models based on Artificial Neural Network (ANN) or on Wavelet Neural Network (WNN) as in (Cuia et al., 2015). In (Huamani et al., 2011), the followed methodology was based on Fuzzy Inference Systems (FIS). However, (Coulibaly and Baldwin, 2005), (Firat, 2008) and (Kisi et al., 2012) have

discussed extensively neuro-fuzzy hybrid models that have the capability of preserving the learning abilities of ANNs and the reasoning of fuzzy systems. Although a variety of forecasting approaches have been successfully formulated, choosing the proper model to accurately predict river flows still imposes a challenge to hydrologists up to date.

In this paper, the study is carried on the Litani River which rises in the central Bekaa Valley. Its water flow is received at the Qaraoun dam, the largest artificial lake in Lebanon. The main concern was dealing with meteorological and hydrological data suffering from insufficiency and also from certain inaccuracies and sometimes unreliability in the information provided by the gauging stations. Besides that, in the past decades, Litani River experienced many major outlaw actions like: major garbage dumping, direct release of urban sewage water, industrial discharges, lack of riverbed maintenance, infringements, and prohibited diversions (International Resources Group (IRG), 2012). Thus, the river flow prediction model features a highly dynamic and non-linear structure. In addition, It accompanies forecasting errors related to noisiness and non-homogeneity of data. However, during the authors literature review, Fuzzy theory appears to be quite effective for handling these aspects, especially when the inherent physical relationships are not fully understood (Nayak and Sudheer, 2008). In addition, according to (Cheng and Li, 2012), Fuzzy Time Series (FTS) has attracted more interest due to its capabilities of dealing with the uncertainty and the vagueness that are often inherent in real-world data resulting from imprecision in measurements, imperfect sets of observations, or difficulties in acquiring measurements under uncertain circumstances. (Bouchon-Meunier et al., 2008) also claim that fuzzy logic provides an interesting tool in the field of data mining, mainly because of its ability to represent flaw information, which is crucial when databases are complex, large, imprecise and contain heterogeneous data. Therefore, in this study, the proposed Fuzzy inference approach (Luna et al., 2007) is adopted for daily river flow time series modeling.

Indeed, the presented method is based on a Constructive-Fuzzy System Modeling (C-FSM) and it is formed of two steps: First, the model is initialized by applying the Subtractive Clustering (SC) algorithm on the available historical data to determine the initial structure of the system (Huamani et al., 2011). In fact, this procedure had provided an extra tool to divide the heterogeneous data into more homogeneous sub-populations which in turn improves the forecasting accuracy (Asadia et al., 2013). Second, the initial structure is modified and refined based on constructive offline learning where a classical Expectation Maximization (EM) algorithm is used for adjusting the parameters of the model.

Up-to-date, the major concern in Fuzzy modeling, is the identification of the suitable input vector. Traditionally, the family of Auto-Regressive models has been widely used for modeling water resources time-series. The order of these models is typically estimated by examining the plots of the Auto-Correlation Function (ACF), Partial Auto-Correlation Function (PACF) and Cross-Correlation Function (CCF). According to (Sudheer et al., 2002) and (Galavi and Shui, 2012), the statistical parameters ACF, PACF and CCF could be also utilized in Fuzzy modeling. Concerning Litani River, the determination of the number of antecedent rainfall, temperature and river flow values involves the computation of time lags that have a significant influence on the forecasting process.

Once relevant inputs are selected, three models are considered: an Auto-Regressive model (AR), a classical Constructive Fuzzy System Modeling (C-FSM) and a Constructive Fuzzy System coupled with a Moving Average filtering method (C-FSM_MA). The Moving Average (MA) aims to reduce rainfall fluctuation and filter out noise. The filtered rainfall data is then fed into the C-FSM forecasting model. As a matter of fact, this technique has been used extensively in (Vos and Rientjes, 2005) and (Wu et al., 2012) for predicting runoff and precipitation respectively via ANN modeling. The time series model of AR type is developed in this study as a benchmark model since the correlation analysis used to determine its structure was also adopted for the Fuzzy modeling.

The main scope of this paper is not solely comparative. It aims to analyze and discuss stochastic modeling of river flow time series using FIS coupled with traditional correlation analysis. (Huamani et al., 2011) have presumed the normality of the gathered data, and in the course of the C-FSM training process they didn't state a clear approach for choosing the appropriate cluster radius. In this work, and to boost the performance of the C- FSM, the collected data had been brought to near a normal distribution using a suitable transformation. In addition, a calibration phase is introduced before validation to select the suitable cluster radius. Moreover, several scenarios were tested by simulating the streamflow associated with different data processing techniques in order to assess their performance.

This study is a part of a wider research project entitled "Operational Optimization of a Multipurpose Cascade Hydropower-Irrigation System". The resulting forecasting model can help in predicting the daily water stock in Qaraoun lake. In fact, water storage is deeply involved in short-term hydro generation optimal scheduling of cascade hydropower stations. The outcome is a multi-functional tool for optimal operation planning capable of enhancing performance of a multi-purpose reservoir system.

The paper is organized as follows: the next section describes the AR, C-FSM and C-FSM_MA models general structure and the utilized optimization algorithm. In Section three, a case study is presented; in fact, this section is divided into two subsections: the first one describes the selected area for the present study and the collected data; the second subsection exhibits data pre-processing techniques and input selection. Section four addresses applications of models, results and discussions. Afterward, in section five conclusions are drawn.

## 2. Models Description

### 2.1. Auto-Regressive Model (AR)

#### 2.1.1. (AR) structure

An Auto-Regressive model (AR) defines the next random variable in a sequence as an explicit linear function of previous ones within a time frame. The structure of AR model of order $p$ is given in equation (1):

$$y(t) + a_1 y(t-1) + ... + a_p y(t-p) = e(t) \tag{1}$$

where $y(t)$ is the output at time $t$, $a_1, ..., a_p$ are the parameters of the AR model to be estimated from the data, $y(t-1), ..., y(t-p)$ are the previous outputs on which the current output depends and $e(t)$ is the white-noise disturbance.
It is known that, the name "autoregressive" comes from the fact that the output $y(t)$ is regressed on the past values of itself.

#### 2.1.2. Optimization algorithm

There are many ways to estimate the coefficients of (1), such as the ordinary least squares procedure, method of moments, Markov chain - Monte Carlo or Yule–Walker methods. However, in this paper Yule–Walker equations are used to relate the Auto-Regressive model parameters to the Auto-correlation coefficient $\rho$ of the random process $y(t)$.

The values of $a_1, ..., a_p$ are determined by solving the matrix equation (2):

$$\begin{pmatrix} 1 & \rho_1 & \cdots & \rho_{p-1} \\ \rho_1 & 1 & \cdots & \rho_{p-2} \\ \rho_2 & \rho_1 & \cdots & \rho_{p-3} \\ \vdots & \vdots & \ddots & \\ \rho_{p-1} & \rho_{p-2} & \cdots & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{pmatrix} = - \begin{pmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ \vdots \\ \rho_p \end{pmatrix} \tag{2}$$

### 2.2. Constructive Fuzzy System Modeling (C-FSM)

#### 2.2.1. C-FSM Structure

Typically, the Multiple Input Single Output (MISO) model structure based on first order Takagi-Sugeno Fuzzy system is composed of a set of $M$ fuzzy rules. Its representative power is manifested through its capability of describing a highly complex nonlinear system using a small number of simple rules.

Let us denote by $\mathbf{x}^k = [x_1^k \; x_2^k \; ... \; x_p^k] \in \mathbb{R}^p$ the input vector at instant $k$ , $k \in \mathbb{Z}^+ - \{0\}$; $\hat{y}^k$ is the output of the model, for a given $\mathbf{x}^k$.The aim is subdividing the input space into $M$ fuzzy sub-regions and approximating the system in each subdivision by a simple linear model. Each partition is defined by its center $c_i \in \mathbb{R}^p$ and its covariance matrix $V_i \in \mathbb{R}^{p \times p}$, whereas a data point can belong to all partitions with different membership degree $g_i^k$ that lies between 0 and 1, such that the sum of all membership values is equal to 1. Afterward, an IF-THEN rule is set to each sub-region; it is defined in the form:

$R_i$ : **IF** $< \mathbf{x}^k$ belongs to the $i^{th}$ region with a membership degree $g_i^k >$ **THEN**

$$y_i^k = \varphi^k \times \theta_i^T \tag{3}$$

Where $\varphi^k = [1 \; x_1^k \; x_2^k \; ... \; x_p^k] \in \mathbb{R}^{p+1}$, and $\theta_i = [\theta_{i0} \, \theta_{i1} \, ... \, \theta_{ip}] \in \mathbb{R}^{p+1}$ is the coefficients vector (parameter) for the local model (Figure 1). Every input pattern has a membership degree associated to each subregion of the input space and is calculated by the formula:

$$g_i(x^k) = g_i^k = \frac{\alpha_i P\left[i \left| x^k \right.\right]}{\sum_{q=1}^{M} \alpha_q P\left[q \left| x^k \right.\right]} \tag{4}$$

where $\alpha_i$ is a positive parameter that is considered as an indirect measure of the relevance of each rule and satisfies $\sum_{i=1}^{M} \alpha_i = 1$. $P\left[i \left| x^k \right.\right]$ is the conditional probability of activating the $i^{th}$rule given the input vector $\mathbf{x}^k$ and is defined as:

$$P\left[i \left| x^k \right.\right] = \frac{1}{(2\pi)^{p/2} \det(V_i)^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}^k - c_i) V_i^{-1} (\mathbf{x}^k - c_i)^T\right\}$$

Where det(.) is the determinant function.

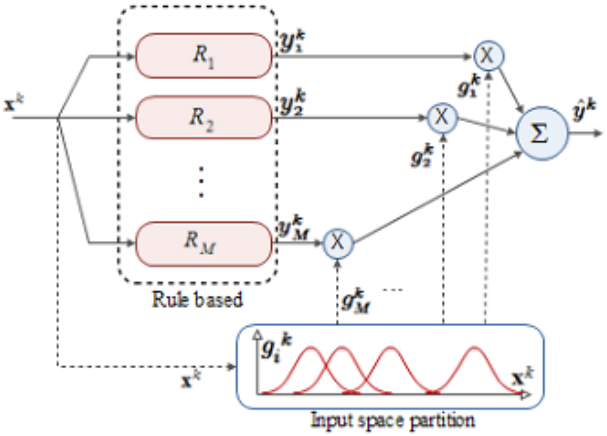


Figure 1. C-FSM general structure with a total of $M$ fuzzy rules (Luna et al., 2007)

The final model output is computed by a non-linear weighted average of the aggregated local outputs and their respective membership degrees. Thus, the estimated output value of the global model for the future time instant $k$ is:

$$\hat{y}^k = \sum_{i=1}^{M} g_i^k y_i^k \tag{5}$$

#### 2.2.2. Optimization Algorithm

The constructive offline learning process for building a FIS model determines automatically the number of fuzzy rules as well as its internal parameters $c_i$, $V_i$, $\theta_i$ and $\alpha_i$ for $i = 1, ..., M$. In fact, the procedure is carried over two stages: model initialization and structure modification stage (Luna et al., 2007).

At stage one, the model is initialized by using the well known Subtractive Clustering algorithm (SC) (Chiu, 1994). Its goal is to determine the initial structure of the fuzzy system that will serve as a starting point for the next stage. The input-output pattern constructed from the available historical data is fed into the SC routine that is available in MATLAB.
The function returns the cluster centers in the matrix C and the vector S which contains the sigma values that specify the range of influence of a cluster center in each of the data dimensions. $r_a$ is a number ranging between 0 and 1 that specifies the cluster center's range of influence, assuming that data falls within a unit hypercube.
Suppose the initial number of rules $M^0$ is the length of the matrix C. Then, the C-FSM structure is initialized as follows:

- $c_i^0 = C_i|_{1...p}$ the first $p$ components of the $i^{th}$ center found by the SC algorithm.
- $V_i^0 = r_a^2 I$, the covariance matrix codifying the spread where $r_a$ is the spread parameter that is used in the SC algorithm and $I$ is the $p \times p$ identical matrix.
- $\theta_i^0 = \left[C_i^{p+1} \ 0 \ldots 0\right]$, $C_i^{p+1}$: last $p+1$ component determined by the SC algorithm.
- $\sigma_i^0 = 1.0$, the initialized standard deviation for each local output $y_i^k$.
- $\alpha_i^0 = 1/M^0$.

Once the model initialization is completed, parameters are re-adjusted based on EM algorithm with the objective of maximizing the log-likelihood $\mathcal{L}$ (6) of the observed values of $y^k$ at each step M of the learning process.

$$\mathcal{L}(D,\Omega) = \sum_{k=1}^{N} \ln\left(\sum_{i=1}^{M} g_i(x^k,\mathbf{C}) \times P(y^k \left| x^k, \theta_i\right.)\right) \quad (6)$$

where $D = \left\{\left(x^k, y^k\right); k = 1, ..., N\right\}$ is the training set, $\Omega$ contains all the model parameters and bold $\mathbf{C}$ contains the centers and the covariance matrix parameters. However, for maximizing $\mathcal{L}$ it is necessary to estimate $h_i^k$: the posterior probability of $x^k$ belong to an active region of the $i^{th}$ local model that is computed for $i = 1, ..., M$ by:

$$h_i^k = \frac{\alpha_i P\left[i \left| x^k\right.\right] P\left[y^k \left| x^k, \theta_i\right.\right]}{\sum_{q=1}^{M} \alpha_q P\left[q \left| x^k\right.\right] P\left[y^k \left| x^k, \theta_q\right.\right]} \quad (7)$$

The conditional probability $P\left[y^k \left| x^k, \theta_i\right.\right]$ is defined as:

$$P\left[y^k \left| x^k, \theta_i\right.\right] = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{\left[y^k - y_i^k\right]^2}{2\sigma_i^2}\right) \quad (8)$$

The variance of the local output $y_i^k$ can be estimated by:

$$\sigma_i^2 = \left(\sum_{k=1}^{N} h_i^k \left[y^k - y_i^k\right]^2\right) / \sum_{k=1}^{N} h_i^k \quad (9)$$

The EM algorithm for finding the parameters is summarized by:

1. E step: Estimate $h_i^k$ via (7)
2. M step: Maximize (6) and update the model parameters:

$$\alpha_i^{new} = \frac{1}{N}\sum_{k=1}^{N} h_i^k \quad (10)$$

$$c_i^{new} = \left(\sum_{k=1}^{N} h_i^k \mathbf{x}^k\right) / \sum_{k=1}^{N} h_i^k \quad (11)$$

$$V_i^{new} = \left[\sum_{k=1}^{N} h_i^k (\mathbf{x}^k - c_i)^T (x^k - c_i)\right] / \sum_{k=1}^{N} h_i^k \quad (12)$$

for $i = 1, ..., M$. An optimal solution for $\theta_i$ is obtained by solving the equation:

$$\sum_{k=1}^{N} \frac{h_i^k}{\sigma_i^2}\left(y^k - \varphi^k \times \theta_i^{new}\right).\varphi^k = 0 \quad (13)$$

After adjusting the parameters, $\mathcal{L}(D,\Omega)$ is recalculated and saved as $\mathcal{L}_{new}(D,\Omega)$.
3. Convergence: Stop the process if:

$$\mathcal{L}_{new}(D,\Omega) - \mathcal{L}_{old}(D,\Omega) < \varepsilon$$

else return to step 1.

### *2.3. Constructive Fuzzy System Modeling coupled with Moving Average (C-FSM_MA)*

To enhance the performance of the forecasting model, the C-FSM_MA adopts the C-FSM structure, whereas its inputs are fed by treated data obtained through the use of the (MA) filter presented in subsection 3.2.

## 3. Case Study

### *3.1. Study Area and Collected Data*

The Litani is the longest river in Lebanon reaching a length of 170 Km. Its watershed covers an area of 2160 $Km^2$ and it is fed by an average level of rainfall around 764 millions of cubic meters (Litani River Authority (LRA), 2016) . It rises near the ancient city of Baalbeck in the central Bekaa Valley, 85 Km east of the Capital Beirut. It flows southward for 100 Km or so before bending sharply toward the west, entering the Mediterranean at Kasmieh just north of the city of Tyre. In the late 1950s, a major development on the Litani River involved constructing the artificial Qaraoun lake (Figure 2) and its associated structures: hydropower plants and irrigation systems. The power plants, located in the vicinity of the river, generate around 190 MW of electric power (10% of the total power production of Lebanon). In the irrigation sector, the use of the Litani River is projected to increase from around 5000

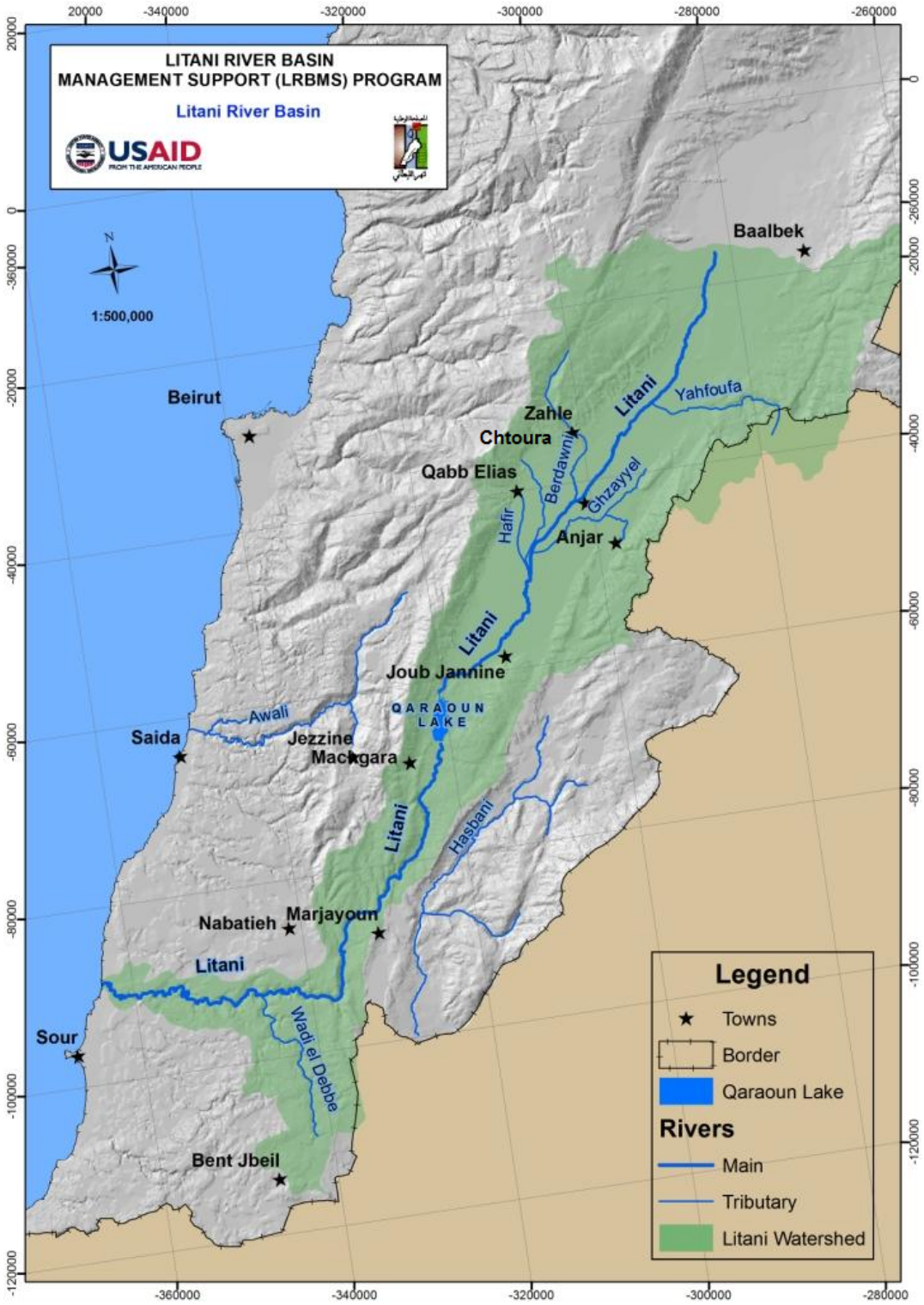


Figure 2. Litaini River Basin (Source: Litaini River Basin Management Support Program - USAID)

ha to more than 21000 ha in the near future (Ramadan et al., 2012).

Due to urbanization and industrialization, the Litani River basin is today experiencing increasing water demands, groundwater over-exploitation, and extensive pollution. As previously mentioned, a walk along the riverside shows: Extensive garbage dumping, direct release of urban sewage water, agricultural run-off, uncontrolled industrial discharges, lack of riverbed maintenance, infringements and prohibited diversions (International Resources Group (IRG), 2012). All these activities are often illegitimate but there are rarely available possibilities for water users to behave differently.

Litani River catchment receives annually 500-600 mm of rainfall (Verner et al., 2013). The peak of rainy season is between December and April, where 75 percent of the rain occurs (Figure 3). Average Temperatures range between $9^oC$ in the winter to $27^oC$ in the summer.

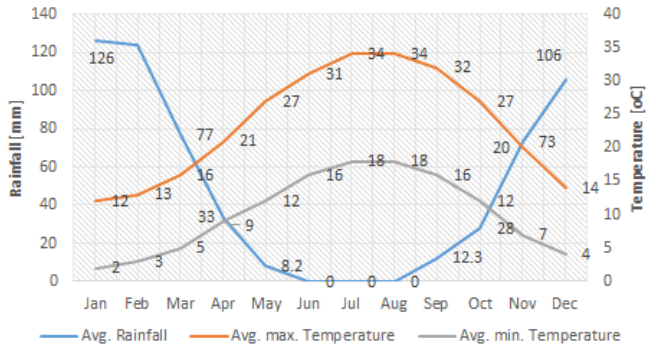


Figure 3. Average rainfall and average max./min temperature (Source:www.worldweatheronline.com)

Long-term hydrological and meteorological data are essential for investigating the river flow regime. For the Litani basin, such data are either incomplete, missing or not available due to civil war and other logistic constraints. In fact, the available data covers only 4 years (June 2009 - December 2013) of rainfall, temperature and river flow daily measurements, retrieved from two distinct places: Machghara weather station and Joub-Jannine hydrological gauging station (Table 1).

Table 1. Machghara and Joub-Jannine stations

| Name Station | Location Latitude | Longitude | Measurements | Duration Daily basis |
|---|---|---|---|---|
| Machghara | 33.5253 | 35.6468 | Rainfall, Temperature | 2009-2013 |
| Joub-Jannine | 33.3821 | 35.4648 | River flow | 2009-2013 |

During the author's visit to Qaraoun dam, an interesting piece of information was revealed: the director in charge claims that the Joub-Jannine streamgage station is not fully automated which may result in frequent gaps and data inconsistency. However, to cover up the gaps, the operators of Joub-Jannine station (upstream) acquire river flow data from the Qaraoun reservoir (downstream) 5-6 Km away. This matter introduces non-homogeneity into data series that was confirmed using Pettitt and Von Neumann homogeneity tests. Thus, besides the uncertainties associated with extreme events (meteorological, hydrological and illegal activities), numerous data limitations affect the accuracy of the results addressed in the paper, including insufficient data and inconsistency due to the fact that some measurements were taken from different sources. All these factors suggest a river system with high variability. Figure 4 shows the mean monthly river flow values with the day-to-day variability at every month. High streamflow, with high variability, occurs in the wet season for the period starting January to April and during December, with a peak flow in February, whereas the river is almost dry from July till October.

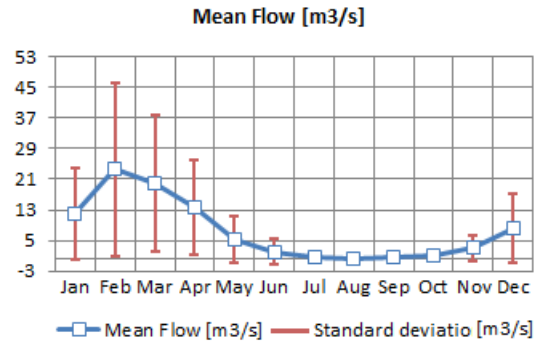


Figure 4. Average and standard deviations of the monthly flow - Litani River

Based on available data, Litani River is characterized by a strong seasonal pattern: high water flow in winter and spring while a low discharge in summer. Further, it possesses great inter-annual variability (Coefficient of variation: $CV = 1.336 > 1$) with a rather weak flow. According to (Leopold et al., 1995), this is mainly due to the fact that the river follows a pluvial regime.

It was found by (Rushworth et al., 2013) that under different climate conditions, the influence of precipitation on flow variability arises due to several reasons: 1- antecedent ground wetness, 2- time-delay in rainfall caused by spatial separation, 3- snow accumulation and melt. Therefore, the rainfall is not the only term that induces variation in the streamflow. In fact, by calculating the coefficient of determination $R^2$ of the available data, only 4.2% of the variation in streamflow is explained by the variation of rainfall.

Figure 5 depicts streamflow, rainfall, and temperature of the entire data set. The following can be noticed: At the middle of the wet season (around January), the fast responding "runoff" causes a more instantaneous response of streamflow to rainfall. In fact, surface runoff accounts for much of the flow during prolonged rainy periods, which wasn't the case at early days of the season. Fast runoff arises when antecedent soil moisture increases to a level where rainfall can move faster near the soil surface without being absorbed. It can result in a rapid increase in flow over a short time period. During the rainy season, runoff

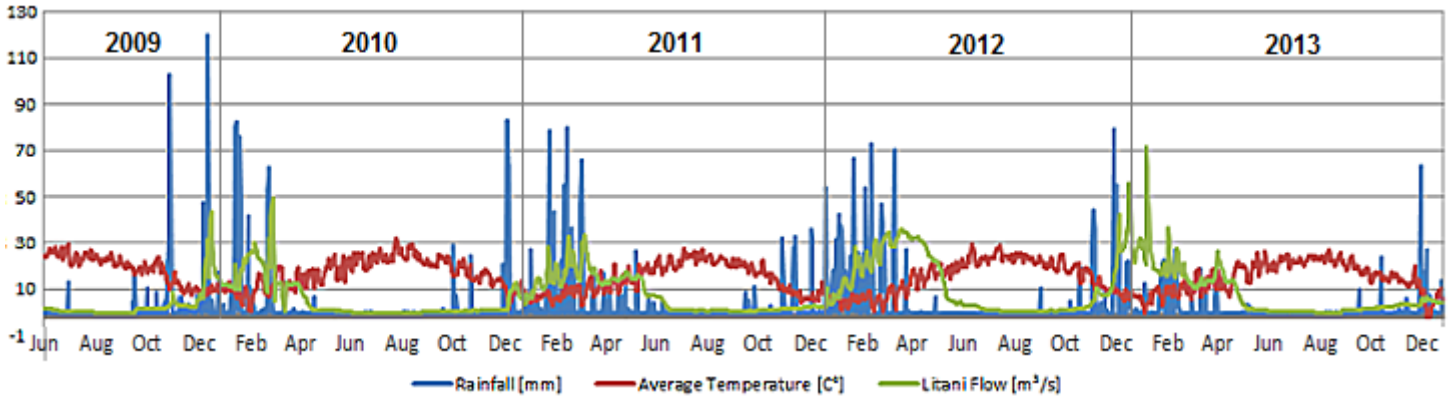


Figure 5. Daily river flow ($m^3/s$), temperature ($^0C$) and rainfall ($mm$) starting 01 Jun 2009 till 31 Dec 2013

in Litani River catchments is one of the most important drivers of variation in flow levels (Rushworth et al., 2013). It is affected by physical factors including soil and subsurface composition, surrounding land usage, evaporation, and transpiration. However, at the early spring months, as the weather gets warmer and the rainfall starts to tapper off, the snowmelt becomes the main driver of the Litani River. Starting April, the river begins to exhibit a decrease in flow and it continues in this manner until it dries around June month when all the accumulated snow at the mountains tops melts off.

In what follows, we proceed by utilizing flexible statistical methods with the aim of constructing a framework that allows us to approximate the flow generating processes with an attempt to identify rainfall-flow, temperature-flow and flow-flow (present-previous flow) relationships.

### *3.2. Data Preprocessing and Input Selection*

The available weather and streamflow measurements, corresponding to the continuous period starting from June 2009 to December 2013, were split into two subsets: a training data set composed of all data preceding January 2013 and a testing set formed of the remaining data.

#### *3.2.1. Standardization/Normalization*

According to (Firat, 2008), (Luna et al., 2007) and (Wu et al., 2012) standardization is crucial in the improvement of both Fuzzy and Auto-Regressive models. In fact, all series presented in this paper: rainfall ($P$), temperature ($T$) and river flow ($Q$) have periodic and seasonal components. They were removed by standardizing the original data through the following transformation:

$$z_m^k = \frac{y_m^k - \mu(m)}{\sigma(m)} \tag{14}$$

where $z_m^k$ is the stationary version of the time series $y^k$ at instant $k$, $\mu(m)$ is the monthly average value and $\sigma(m)$ is the monthly standard deviations.

For the standardization process of the river data, average flow was considered with the monthly values presented in Figure 4 along with their standard deviation.

Moreover, in the course of Fuzzy System Modeling (FSM), the model is initialized using Subtractive Clustering (SC) with spread radius $r_a \in [0, 1]$, Thus, it is necessary to re-scale or normalize the trained data set within a unit hypercube, using the formula:

$$Z_{norm}^k = \frac{z^k - z_{\min}}{z_{\max} - z_{\min}} \tag{15}$$

where $Z_{norm}^k$ is the normalized data at time $k$, $z^k$ is the observed value, $z_{\min}$ and $z_{\max}$ are the minimum and maximum in the data set.

#### *3.2.2. Data Filtering via Moving Average (MA)*

MA filters data by replacing each data point with the average of the neighboring $k$ data points, where $k$ is the size of the memory window. The method is based on the idea that any large irregular component at any point in time will exert a smaller effect if we average the point with its immediate neighbors (Newbold et al., 2003). The equally weighted MA is the most commonly used method, where each value of the data carries the same weight in the data filtering process. The $k$-term unweighted moving average $y_t^*$ can be calculated by:

$$y_t^* = \frac{1}{k}\sum_{i=0}^{k-1} y_{t-i} \tag{16}$$

where $t = k; ..., N$.

#### *3.2.3. Correlation analysis*

One of the most important steps in the forecasting model development process is the determination of significant input variables. The employed statistical approach in this study was suggested by (Sudheer et al., 2002) to identify the appropriate input vector. The method is based on

the heuristic that the possible influencing variables, related to different time lags, can be identified through correlation analysis (Shanmuganathan and Samarasinghe, 2016). Basically, Cross Correlation and Partial Auto-Correlation between the variables are utilized.

Using available training data, the PACF suggests a significant correlation at 95% confidence level up to 6 days of river flow lag (Figure 6). One may notice that lag 6 shows better significance than lags 4 and 5. This anomaly is closely related to the limited data, since lag 6 had dropped below threshold once the PACF is carried on the whole data set (training and testing). During the transformation

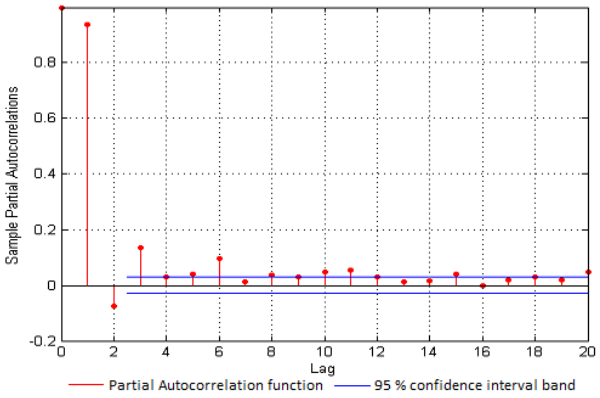


Figure 6. Partial Auto-Correlation function (PACF) of Litani flow

of rainfall into streamflow, the rainfall input to the system goes through two operators: i -"translation" in time; and ii- "attenuation" due to the storage characteristics of the watershed (Chow et al., 1988). The sophistication and complexity of these two operations may explain the weak Cross Correlation between rainfall and streamflow (Figure 7). In order to enhance the similarity between rainfall and streamflow, (Wu et al. (2012)-Wu et al. (2009)) explored the efficiency of various data pre-processing methods in improving the input-output mapping of the ANN model by filtering raw data. One of the used techniques is the Moving Average (MA).

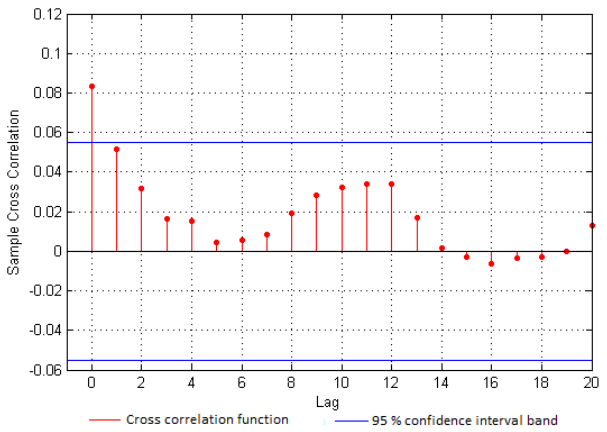


Figure 7. Cross-Correlation Function (CCF) between unfiltered rainfall and Litani flow

In this paper, the MA operation entails the window size $k$ in Eq. (16) to filter the raw rainfall data. A suitable $k$ was found by a systematic increase of $k$ from 1 to 12, where at every step, the filtered data is cross-correlated with the river flow data. The targeted value of $k$ corresponds to the optimal zero-lag Cross-Correlation. Physically, it is known that, Cross-Correlation measures the similarity between two signals. Thus $k$ was chosen in a way that reveals the best similarity between rainfall and streamflow.

The plot in Figure 8 shows that the best zero lag correlation occurs at a window size 12. Thus MA(12) is adopted for the filtering process. It is clear that the filtered rainfall

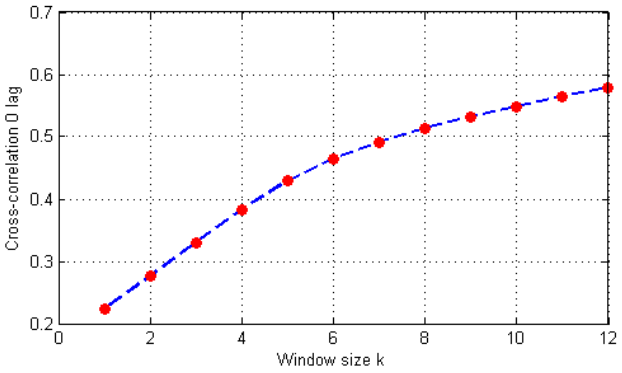


Figure 8. Moving Average (MA) window size $k$ versus zero lag Cross-Correlation

data exhibits better correlation than the unfiltered one when cross correlated with the river-flow (Figures 7, 9). We note here that wider window size was not considered

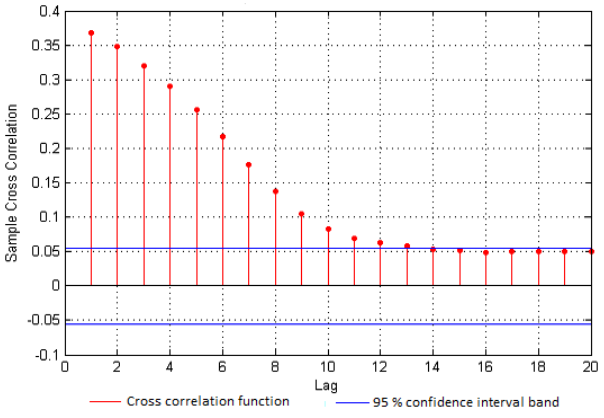


Figure 9. Cross-Correlation Function (CCF) between filtered rainfall and Litani flow

since the improvement in Cross-Correlation was negligible. Figure 10 exhibits the enhanced similarity between the river flow and the filtered rainfall data.

Furthermore, the temperature ($T$) and the river-flow ($Q$) were also cross-correlated and the result showed a negative correlation up to lag 20 which can be interpreted as ($T$) varies in opposite sense with ($Q$). Therefore, ($T$) is also considered as an input in the suggested models.

### *3.2.4. Data Transformation*

According to (Aqil et al., 2007) , networks trained on transformed data attain better performance. In this pa-

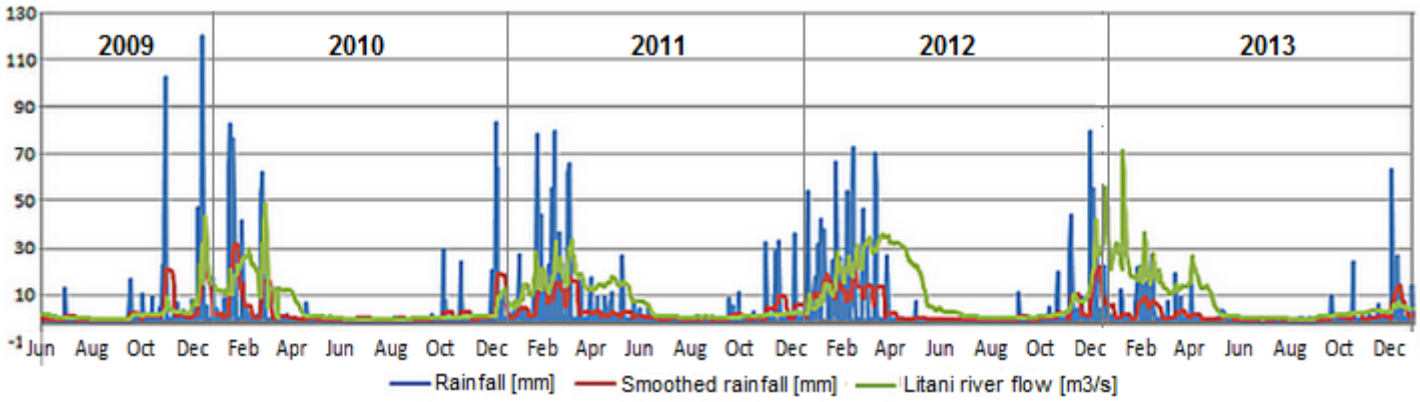


Figure 10. Daily river flow ($m^3/s$), rainfall ($mm$) and filtered rainfall ($mm$) starting 01 Jun. 2009 till 31 Dec. 2013.

per, a log transformation has been considered to bring the observed data as much as possible to resemble a normal distribution. The log transformation is performed on each input and output variable independently, using the following equation:

$$Y = a \log_{10}(X + b) \tag{17}$$

The forecasted results are then back-transformed using the inverse transformation:

$$X = 10^{Y/a} - b \tag{18}$$

Where $a$ and $b$ are arbitrary constants.
The coefficients $a$ and $b$ of (17) are obtained on trial-and error basis, until the data follow a normal distribution. For $a = 0.5$ and $b = 1$, the descriptive statistics of the entire data is shown in Table 2.

Table 2. Statistical properties of raw and logarithmic transformed daily data

| Dataset | Entire (Observed) | Entire (transformed) | Training (transformed) | Testing (transformed) |
|---|---|---|---|---|
| Statistics | (2009-2013) | (2009-2013) | (2009-2012) | (2013) |
| **Daily Rainfall** $[mm]$ | | | | |
| Mean | 2.993 | 0.092 | 0.098 | 0.070 |
| St. Deviation | 10.797 | 0.213 | 0.223 | 0.170 |
| Skewness | 5.262 | 2.485 | 2.405 | 2.653 |
| Kurtosis | 32.578 | 5.171 | 4.681 | 6.143 |
| CV | 3.608 | 2.320 | 2.282 | 2.417 |
| **Smoothed Daily Rainfall** $[mm]$ | | | | |
| Mean | 2.985 | 0.175 | 0.189 | 0.125 |
| St. Deviation | 5.268 | 0.211 | 0.221 | 0.157 |
| Skewness | 2.253 | 1.016 | 0.909 | 1.222 |
| Kurtosis | 4.985 | -0.288 | -0.581 | 0.446 |
| CV | 1.765 | 1.203 | 1.170 | 1.263 |
| **Daily River flow** $[m^3/s]$ | | | | |
| Mean | 7.463 | 0.318 | 0.314 | 0.329 |
| St. Deviation | 9.973 | 0.253 | 0.255 | 0.244 |
| Skewness | 1.670 | 0.402 | 0.428 | 0.312 |
| Kurtosis | 2.926 | -1.283 | -1.296 | -1.215 |
| CV | 1.336 | 0.796 | 0.812 | 0.741 |

It can be noticed from Table 2 that the statistical indicators: standard deviation, skewness and kurtosis show high values for observed data. After the logarithmic transformation, these indicators were reduced significantly. However, regarding the temperature, the skewness and kurtosis were relatively small. Thus, in this study, there is no need to consider data transformation for the temperature.

### *3.2.5. Input Selection*

This paper aims modeling river flow process by AR and FSM models by using recorded rainfall, temperature and streamflow data. Based on the graphical interpretation of PACF and CCF, several input combinations of river flow, rainfall and temperature were examined in the modeling process. The input pattern considers both past and present precipitations $(..., P_{t-2}, P_{t-1}, P_t)$ and temperatures $(..., T_{t-2}, T_{t-1}, T_t)$ but only past stream data $(...Q_{t-3}, Q_{t-2}, Q_{t-1})$ for the river flow. The output corresponds to the present river flow $(Q_t)$, where the subscript $t$ represents the time step. As a consequence, different input combinations of $Q$, $P$ and $T$ data were constructed and listed in Table 3.

Table 3. Model Structure: input-output configuration

| Model | Input structure | Output |
|---|---|---|
| AR | $Q_{t-1}, Q_{t-2}, ..., Q_{t-7}$ | $Q_t$ |
| C-FSM | $Q_{t-1}, Q_{t-2}, ..., Q_{t-7}, P_t, P_{t-1}, T_t, T_{t-1}$ | $Q_t$ |
| C-FSM_MA | $Q_{t-1}, Q_{t-2}, ..., Q_{t-7}, P_t, P_{t-1}, ..., P_{t-8}, T_t, T_{t-1}$ | $Q_t$ |

Another major input that needs to be identified in the Fuzzy modeling is the cluster radius. We recall that the radius specifies the range of influence of the cluster center on each input-output point. Knowing that the cluster radius falls within the unit hypercube, a smaller cluster radius yields an increase in clusters and thus a greater number of rules which will increase the model's complexity. However (Velasquez and Palade, 2013) suggests that, the best value for a given radius is usually between 0.2 and 0.5, so the clustering radius is identified through a trial and error procedure by varying the cluster radius from 0.2 to 0.5 with an increment of 0.01 to get the best performance during the calibration phase.

## 4. Results and Discussions

### 4.1. Performance Metrics

Models performance and testing were conducted considering the Root Mean Square Error (RMSE-$[m^3/s]$), Mean Absolute Error (MAE-$[m^3/s]$) and the Mass Curve Coefficient (E) defined by the following equations:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}^k - y^k)^2}{n}} \tag{19}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}^k - y^k\right| \tag{20}$$

$$E = \frac{\sum_{i=1}^{n}(y^k - \overline{y})^2 - \sum_{i=1}^{n}(y^k - \hat{y}^k)^2}{\sum_{i=1}^{n}(y^k - \overline{y})^2} \tag{21}$$

### 4.2. Some Tweaks

For the sake of being fair with all Fuzzy models, a calibration phase was considered and performed on one month data (December 2012). Since FSM is sensitive to the number of clusters; the best radius $r_a$, that corresponds to the optimal efficiency E for 1 day lead forecast, is achieved by varying it from 0.2 to 0.5 with an increment of 0.01 and limiting the number of clusters between 2-8 to avoid over fitting (Nayak and Sudheer, 2008).

Table 4 displays different scenarios with the utilized data processing method and the optimization algorithm. Further it presents, in case of Fuzzy modeling, the used cluster radius (obtained during calibration) and whether the model is coupled with an Adding Operator (Luna et al., 2007).

Table 4. AR, C-FSM and C-FSM_MA Models

| Scenario | | Data Pre-processing | Optimization Algorithm | | |
|---|---|---|---|---|---|
| AR(7) | | Std | Yule-Walker | | |
| | | | Expectation-Maximization | | |
| Group | | | SC→ $r_a$ | Cluster no | AO |
| 1 | C-FSM 1 | Std/Norm | 0.36 | 7 | No |
| | C-FSM_MA 1 | MA/ Std/ Norm | 0.47 | 4 | No |
| 2 | C-FSM 2 | Std/Norm | 0.36 | 7 | Yes |
| | C-FSM_MA 2 | MA/ Std/ Norm | 0.47 | 4 | Yes |
| 3 | C-FSM 3 | Dtrans / Std/ Norm | 0.37 | 7 | No |
| | C-FSM_MA 3 | MA/ Dtrans/ Std/ Norm | 0.36 | 8 | No |
| 4 | C-FSM 4 | Dtrans / Std/ Norm | 0.37 | 7 | Yes |
| | C-FSM_MA 4 | MA/ Dtrans/ Std/ Norm | 0.36 | 8 | Yes |

Std: Standardization, Norm: Normalization, Dtrans: Data Transformation, MA: Moving Average, SC: Subtractive Clustering, AO: Adding Operator

Regarding normality, (Huamani et al., 2011) asserted that the data have to be normally distributed before the model coefficients can be estimated, while (Mehmmet, 2009) claimed that the normality assumption is not restrictive and good results can be obtained by using real world observations directly. In the current application, this issue is investigated by comparing performance of the models developed on transformed (into the normal domain) and non-transformed data.

With respect to (George and Mallery, 2010), skewness and kurtosis values lying between -2 and +2 are considered acceptable in order to prove normal univariate distribution. Thus, after the transformation (Table 2), the smoothed rainfall and streamflow data satisfy the claim of (George and Mallery, 2010) concerning normality. On the other hand, the raw rainfall data was pushed as much as possible to match a normal distribution.

### 4.3. Obtained results

Table 5 shows the performance metrics of 12 prediction horizons for the 9 scenarios carried in the real world (i.e. results were restored to the original space). After fitting the historical flow data to the benchmark AR model of order 7, it shows, for all lead days, the poorest forecasting among the other models. This is due to the fact that AR models are unlikely able to capture any nonlinear dependency. Whereas, the performance of the Fuzzy models accompanied with different pre-processing techniques were more useful for detecting nonlinearities in the streamflow.

Furthermore, due to estimation errors of the previous steps that are fed into the input pattern for the next step ahead, one can notice from observing the performance indices of all scenarios a decreasing trend in the Mass Curve Coefficient (E) and an increase in the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE). Based on the obtained results, the models efficiency (E) in explaining the hydrological process range between 54.4% and 89%.

#### 4.3.1. Effect of data transformation on model performance

For 12 days lead, the AR(7) model gave a RMSE of 6.668 $m^3/s$, the C-FSM 1 model with un-transformed inputs gave RMSE of 4.924 $m^3/s$, i.e. a reduction of 26.15% versus AR. However, the C-FSM 3 model with transformed inputs reduces the RMSE by 30.31% with an improvement of 4.16% more than C-FSM 1 model. In a way, this result supports the claim of (Mehmmet, 2009) that using un-transformed data can still provide good results. In general, results presented in Table 5 showed that the C-FSM 3 model whose inputs are transformed are more accurate (in terms of the Mass Curve Coefficient E) than C-FSM 1, where it emerges as a better performer for most lead days.

On the other hand, C-FSM 4 did not exhibit a clear better performance than C-FSM 2, neither did C-FSM 2. Apparently, the adding operator didn't work well for both models with transformed and un-transformed inputs.

#### 4.3.2. Impact of (MA) filter on model performance

Upon using the (MA) filter, a significant observation was obtained. The coefficient of determination $R^2$ of the sub-series joining the filtered rainfall and the river flow indicates a value of 0.2786. That is, 27.86% of the variability in river flow is explained by that of the filtered

Table 5. Performance measures of forecasting daily river flow for a horizon $h$ varying from 1 to 12

| Scenario | Performance index | Horizon [h] | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| AR(7) | RMSE $[m^3/s]$ | 3.454 | 3.700 | 4.075 | 4.192 | 5.003 | 5.240 | 4.811 | 4.633 | 4.862 | 5.094 | 5.034 | 6.668 |
| C-FSM 1 | | 3.377 | 4.276 | 4.623 | 5.206 | 4.687 | 4.950 | 4.523 | 5.099 | 4.564 | 4.912 | **4.640** | 4.924 |
| C-FSM_MA 1 | | 3.349 | 3.534 | 3.868 | 3.763 | 4.987 | 4.491 | 4.637 | 4.373 | **4.640** | **4.754** | 4.827 | 5.774 |
| C-FSM 2 | | 3.346 | 4.283 | 4.615 | 5.184 | **4.510** | 4.985 | 4.777 | 5.040 | 4.708 | 5.060 | 4.822 | 5.125 |
| C-FSM_MA 2 | | **3.301** | 3.486 | 3.816 | 3.698 | 4.899 | 4.625 | 4.213 | **4.181** | 4.644 | 4.913 | 4.948 | 4.960 |
| C-FSM 3 | | 3.324 | **3.413** | 3.717 | 3.613 | 4.941 | 4.183 | 4.455 | 4.471 | 4.995 | 5.063 | 5.119 | 4.647 |
| C-FSM_MA 3 | | 3.344 | 3.496 | 3.823 | 3.722 | 4.891 | 4.351 | 4.448 | 4.281 | 4.807 | 4.885 | 4.979 | 5.127 |
| C-FSM 4 | | 3.352 | 4.178 | 3.691 | 3.527 | 5.020 | 4.371 | 4.831 | 4.585 | 5.360 | 5.297 | 5.354 | 4.715 |
| C-FSM_MA 4 | | 3.348 | 3.433 | **3.632** | **3.474** | 4.885 | **3.757** | **4.203** | 4.203 | 4.807 | 4.798 | 5.054 | **4.35**0 |
| AR(7) | MAE $[m^3/s]$ | 0.951 | 1.203 | 1.458 | 1.524 | 1.824 | 1.957 | 1.939 | 1.961 | 1.916 | 2.090 | 2.046 | 2.695 |
| C-FSM 1 | | 0.924 | 1.260 | 1.438 | 1.743 | 1.759 | 1.786 | 1.743 | 2.160 | 1.873 | 2.048 | **1.745** | **1.801** |
| C-FSM_MA 1 | | 0.917 | 1.156 | 1.369 | 1.403 | 1.751 | 1.760 | 1.804 | 1.826 | 1.821 | 1.880 | 1.887 | 2.412 |
| C-FSM 2 | | 0.928 | 1.291 | 1.409 | 1.762 | 1.727 | 1.833 | 1.923 | 2.213 | 1.970 | 2.240 | 1.950 | 2.136 |
| C-FSM_MA 2 | | 0.907 | 1.138 | 1.347 | 1.379 | 1.767 | 1.764 | **1.658** | 1.722 | 1.804 | 2.027 | 1.954 | 2.227 |
| C-FSM 3 | | 0.842 | **1.031** | **1.241** | **1.285** | **1.715** | 1.540 | 1.700 | 1.715 | 1.818 | 1.965 | 1.887 | 1.943 |
| C-FSM_MA 3 | | 0.884 | 1.110 | 1.326 | 1.368 | 1.734 | 1.673 | 1.753 | 1.746 | 1.812 | 1.900 | 1.867 | 2.168 |
| C-FSM 4 | | **0.841** | 1.229 | 1.289 | 1.301 | 1.783 | 1.620 | 1.834 | 1.799 | 2.053 | 2.209 | 2.198 | 1.973 |
| C-FSM_MA 4 | | 0.898 | 1.122 | 1.264 | 1.311 | 1.788 | **1.482** | 1.725 | **1.705** | **1.800** | **1.831** | 1.922 | 1.870 |
| AR(7) | E | 0.878 | 0.860 | 0.830 | 0.820 | 0.743 | 0.719 | 0.763 | 0.780 | 0.758 | 0.734 | 0.740 | 0.544 |
| C-FSM 1 | | 0.888 | 0.820 | 0.791 | 0.734 | 0.784 | 0.762 | 0.799 | 0.747 | **0.797** | 0.765 | **0.789** | 0.764 |
| C-FSM_MA 1 | | 0.885 | 0.872 | 0.847 | 0.855 | 0.745 | 0.796 | 0.780 | 0.806 | 0.782 | **0.771** | 0.762 | 0.663 |
| C-FSM 2 | | **0.890** | 0.820 | 0.791 | 0.736 | **0.800** | 0.758 | 0.776 | 0.753 | 0.784 | 0.751 | 0.772 | 0.744 |
| C-FSM_MA 2 | | 0.888 | 0.876 | 0.852 | 0.860 | 0.754 | 0.783 | **0.819** | **0.823** | 0.781 | 0.755 | 0.750 | 0.751 |
| C-FSM 3 | | 0.887 | **0.881** | 0.858 | 0.866 | 0.750 | 0.821 | 0.797 | 0.795 | 0.744 | 0.737 | 0.731 | 0.779 |
| C-FSM_MA 3 | | 0.885 | 0.875 | 0.850 | 0.858 | 0.755 | 0.806 | 0.797 | 0.812 | 0.763 | 0.755 | 0.746 | 0.731 |
| C-FSM 4 | | 0.885 | 0.821 | 0.860 | 0.873 | 0.742 | 0.804 | 0.761 | 0.785 | 0.706 | 0.712 | 0.706 | 0.772 |
| C-FSM_MA 4 | | 0.885 | 0.879 | **0.865** | **0.876** | 0.755 | **0.855** | **0.819** | 0.819 | 0.763 | 0.764 | 0.738 | **0.806** |

rainfall. Therefore, the explained variability has increased from 4.18% to 27.86% when using filtered instead of raw rainfall data. This can be interpreted by the fact that, Moving Average contains within a "memory" that has the ability to record, to a certain extent, the variation caused by snow melt and antecedent ground wetness resulted from previous precipitations. Thus (MA) didn't just remove the noise but it has improved the explained variance by more than 27%.

The impact of the (MA) filter on the performance of the C-FSM model is described as follows: each of the three models C-FSM_MA 1, C-FSM_MA 2 and C-FSM_MA 4 exhibits a noticeable prediction efficiency in many lead days in terms of RMSE, MAE and E compared with C-FSM 1, C-FSM 2 and C-FSM 4. But the remarkable performance is achieved by C-FSM_MA 4 that shows almost the lowest RMSE, the lowest MAE and the highest E. Figure 11 shows the time series plot for 1, 4, 9 and 12 days ahead forecast associated with C-FSM_MA 4 model. During the beginning and the end of the wet season (October, November, December and March, April, June), the flow variability is low and the model shows a noticeable fit with the actual flow. However, it wasn't the case during January and February months.

Furthermore, the actual flow is characterized with a very sharp spike (12 Jan. 2013). This can be interpreted as an anomaly in the observation due to inaccurate measurements and can't be considered as a flood for two main reasons:

- The accumulated rainfall, a week before the spike occurrence date, was only 21 mm, and
- The average temperature during this week was below 6.5 degrees Celsius.

These two reasons are not enough to produce a sudden elevation in the river flow from 20 $m^3/s$ to 71 $m^3/s$ based on previous observations (Figure 5).

Besides inaccurate measurements, the river flow forecast is disrupted by vast sources of noise due to: illegal activities previously mentioned (International Resources Group (IRG), 2012), some meteorological conditions (wind, evaporation, irradiance, etc.) and urbanization. All these factors distort the accuracy of the river flow model and cause a decrease in the forecasting precision. To reduce the noise effect, one can consider different noise filters. However, for the mentioned sources of distortion, quantitative data is not available. Thus, the attention is turned to the noise existing within the rainfall data and (MA) filter. In this case, the Noise to Signal ratio (Jayawardena and Gurung, 2000) was calculated for the two time series (12 days ahead) C-FSM 4 and the denoised one C-FSM_MA 4. The obtained respective values were 0.48 and 0.44. Hence, the MA filter applied to rainfall has reduced the noise in the streamflow time series by 8.33% . This percentage is an acceptable value bearing in mind the scarcity of noise sources data. Furthermore, some people claim that filtering may remove noise as well as variability. Therefore, it might not be a

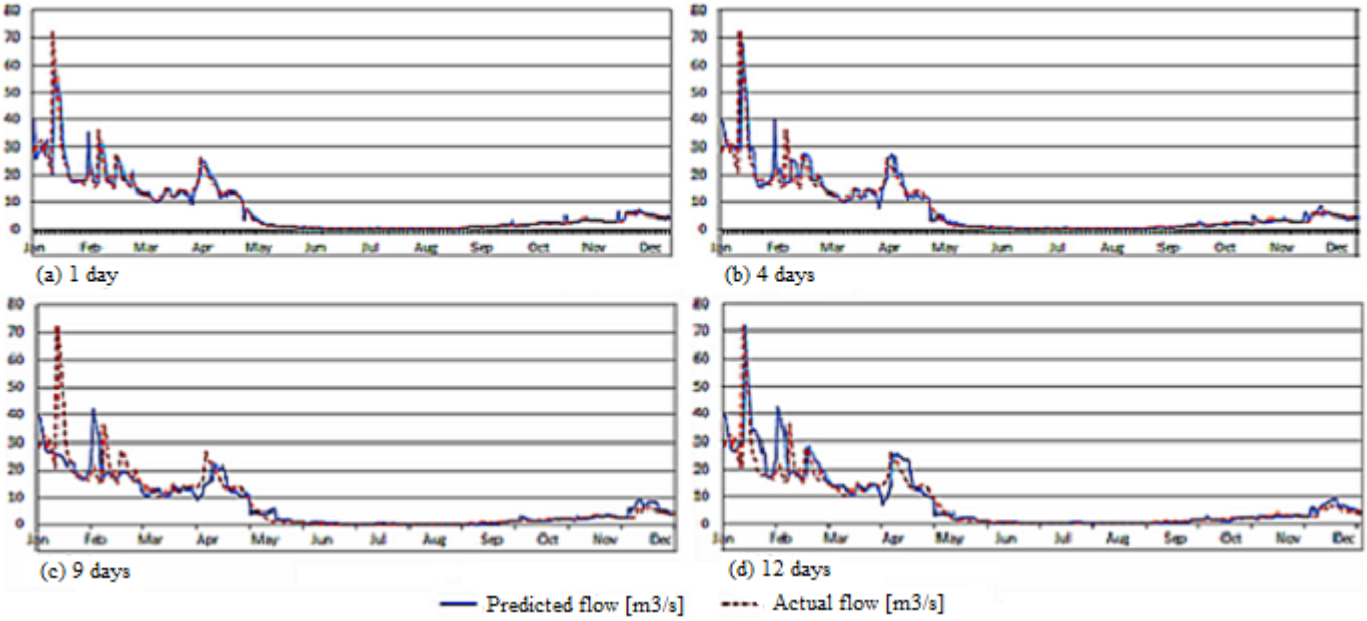


Figure 11. C-FSM_MA 4 river flow estimates (a) 1 day (b) 4 days (c) 9 days (d) 12 days ahead along with the observed flow for Litani River over the testing period (1 January - 31 December, 2013)

good choice for daily streamflow forecasts. In fact, this issue is explained and discussed in the next paragraph.

Although, the performance of C-FSM_MA 4 model was more than satisfactory with an efficiency reaching 80.6% for the 12 days ahead forecast. The model was also able to explain 84.27% of the actual river flow variability. The C-FSM_MA 4 model managed to reproduce the day-to-day variability almost within naturally occurring ranges by taking the "memory" advantage of (MA) filter. Whereas, the C-FSM 4 model that was fed with unfiltered rainfall could capture 79.52 % only.

Further, the nonlinearity of streamflow processes is also investigated. (Brock et al., 1996) introduced a test for the existence of nonlinearity in streamflow processes. It is found that the shorter the time-scale, the stronger the nonlinearity. All annual series are linear, whereas all daily streamflow processes are strongly nonlinear. Looking backward to the linear benchmark model, the Auto-Regressive time series forecast was correlated with the original river flow. It revealed that, for 12 days ahead forecast, the coefficient of correlation is 0.86, which means that nonlinear identification was difficult and the AR, as expected, manifests not much of accurate results. However, regarding C-FSM_MA 4 model, the coefficient of correlation between the observed and the forecasted flows for 12 days lead is equal to 0.92. Thus, this model is more competent in capturing the nonlinearity in river flows at different lead days.

Figure 12 shows the scatter plot of both the observed and the predicted flows obtained by using the C-FSM_MA 4 model on the testing period for 12 days lead. The line y = x represents the perfect fit case when the predicted and the observed river flows are equal. In fact, the reader can notice that, along the line y = x, a tight dispersion for the low flows and a wide one for the high flows (within the circle). Thus, based on the data distribution for high and low flows, the forecasting model showed a good prediction accuracy for the low values of the flow but it was unable to maintain the same accuracy for the high values.

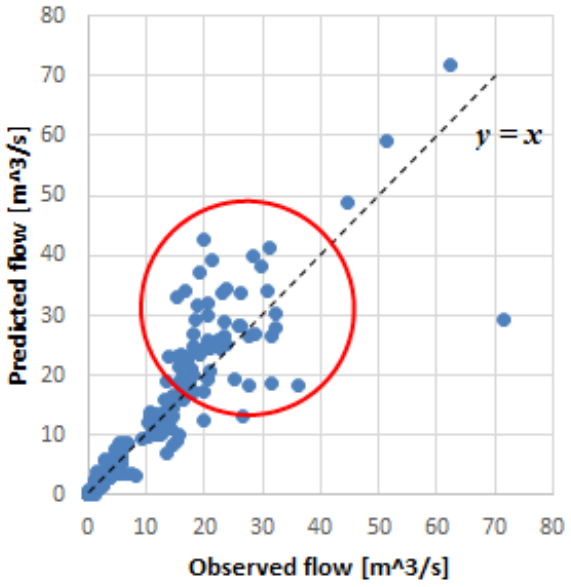


Figure 12. Observed versus predicted river flow for the testing period: 01 Jan. till 31 Dec. 2013

In general, models performance in reproducing and inferring river flow for the testing year were more than satisfactory, given the limitations descending especially from the quality and quantity of the historical observations. If

somehow, an advanced data acquisition system was installed on the river's site, it will have the ability to obtain more accurate and reliable meteorological and hydrological measurements. Thus by sweeping off uncertainties related to missing or inaccurate observations, the models would have delivered even better results. Further, reliable and longer climate and discharge measurements would have allowed a proper training and testing of the model performance. The data scarcity did not allow to account for other sources of uncertainty, such as factors related to climate change and urbanization. However, the C-FSM models proved to be accurate enough to provide plausible results and a reasonable agreement with the observed streamflow. Thus, they were robust enough to be used in a situation where data possess a certain level of heterogeneity.

## 5. Conclusion

In this paper, a comparative study was presented between Auto-Regressive (AR), Constructive Fuzzy System Modeling (C-FSM) and the Constructive Fuzzy System Modeling coupled with Moving Average (C-FSM_MA) methods for multi-step-ahead daily river flow time series forecasting. For achieving this objective, the Litani River - Lebanon was selected as a case study. The suggested models with different inputs variables were trained and tested. Then the results were compared and evaluated using three statistical indicators (RMSE, MAE, and E). Despite the scarcity, heterogeneity and non-normality of meteorological-hydrological data aside with uncertainties inherited from illegal activities reported along the river, due to factors like urbanization and industrialization, the outcomes of the C-FSM and C-FSM_MA models came very satisfactory. Furthermore, the Moving Average filter had provided a supportive tool during fuzzy modeling. It didn't just reduce the noise inherent within rainfall data but it has also preserved the streamflow variability due to rainfall. Overall, the analysis presented in this study provides that, a variant of the C-FSM_MA model had shown a better accuracy over the rest of the models in the river flow forecasting.

Although, the available data suffers from different types of drawbacks, the data driven models based on constructive fuzzy system modeling was successfully applied to establish river flow with plausible performance. These results motivate the authors to adopt the methods suggested in this paper for generating future streamflow scenarios as a part of short-term hydropower operation scheduling. The time series forecast will help in finding the optimal operation policy at different stages through right discharge decisions.

## Acknowledgment

This paper was supported by a Grant from the National Council for Scientific Research - Lebanon (CNRS-L) as a part of a project entitled "Operational Optimization of a Multipurpose Hydropower-Irrigation System". Further, the authors would like to acknowledge the people in charge of the Water Resources Department - Litani River Authority (LRA) and the Agricultural Research Institute of Lebanon (LARI) for providing us with the available hydrological and meteorological data.

Aqil, M., Kita, I., Yano, A., and Nishiyama, S. (2007). A comparative study of artificial neural networks and neuro-fuzzy in continuous modeling of the daily and hourly behaviour of runoff. *Journal of Hydrology*, (337):22–34.

Asadia, S., Shahrabia, J., Abbaszadehb, P., and Tabanmehra, S. (2013). A new hybrid artificial neural networks for rainfallrunoff process modeling. *International Work Conference on Artificial Neural Networks*, (121):470–480.

Bouchon-Meunier, B., Rifqi, M., and Lesot, M. (2008). Similarities in fuzzy data mining: From a cognitive view to real-world applications. *Computational Intelligence: Research Frontiers - Springer*, (5050):349–367.

Box, G. and Jenkins, G. (1970). Time series analysis: forecasting and control. *San Francisco: Holden-Day*.

Brock, W., Dechert, W., J.A., Scheinkman, and LeBaron, B. (1996). A test for independence based on the correlation dimension. *Econ. Rev.*, (15):197–235.

Cheng, Y. and Li, S. (2012). Fuzzy time series forecasting with a probabilistic smoothing hidden markov model. *IEEE Transactions on Fuzzy Systems*, (20):291–304.

Chiu, S. (1994). A cluster estimation method with extension to fuzzy model identification. *World Congress on Computational Intelligence, Proceedings of the Third IEEE Conference*.

Chow, V., Maidment, D., and Mays, L. (1988). Applied hydrology. *McGraw-Hill, New York*.

Coulibaly, P. and Baldwin, C. (2005). Nonstationary hydrological time series forecasting using nonlinear dynamic methods. *Journal of Hydrology*, (307):164–174.

Cuia, Q., Wanga, X., Lib, C., Caia, Y., and Liangd, P. (2015). Improved thomas–fiering and wavelet neural network models for cumulative errors reduction in reservoir inflow forecast. *Journal of Hydro-environment Research*, (In Press).

Dibike, Y. and Solomatine, D. (2001). River flow forecasting using artificial neural networks. *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere*, (26):1–7.

Firat, M. (2008). Comparison of artificial intelligence techniques for river flow forecasting. *Hydrol. Earth Syst. Sci*, (12):123–139.

Galavi, H. and Shui, L. (2012). Neuro-fuzzy modelling and forecasting in water resources. *Scientific Research and Essays*, (7):2112–2121.

George, D. and Mallery, M. (2010). Spss for windows step by step: A simple guide and reference. *Boston: Pearson 10th edition*.

Huamani, L., Ballini, R., Hidalgo, I., Barbosa, P. F., and Francato, A. (2011). Daily reservoir inflow forecasting using fuzzy inference systems. *IEEE International Conference on Fuzzy systems*, (4):201–213.

International Resources Group (IRG) (2012). Litani river basin management plan. *USAID*.

Jayawardena, A. and Gurung, A. (2000). Noise reduction and prediction of hydrometeorological time series: dynamical systems approach vs. stochastic approach. *Journal of Hydrology*, (228):242–264.

Kisi, O., Shiri, J., and Nikufa, B. (2012). Forecasting daily lake levels using artificial intelligence approaches. *Computers and Geosciences*, (4):169–180.

Leopold, L., Wolman, M., and Miller, J. (1995). Fluvial processes in geomorphology. *New York: Dover Publications*.

Litani River Authority (LRA) (2016). The characteristics of the litani river. http://www.litani.gov.lb/.

Luna, I., Soares, S., and Ballini, R. (2007). A constructive-fuzzy system modeling for time series forecasting. *International Joint Conference on Neural Networks*.

Mehmmet, O. (2009). Comparison of fuzzy inference systems for streamflow prediction. *Hydrological Sciences*, (54):261–273.

Nayak, P. and Sudheer, K. (2008). Fuzzy model identification based on cluster estimation for reservoir inflow forecasting. *Hydrological Processes*, (22):827–841.

Newbold, P., Carlson, W., and Thorne, B. (2003). Statistics for business and economics. *Fifth Version, Prentice Hall, Upper Saddle River, NJ*.

Porporato, A. and Ridolf, L. (2001). Multivariate nonlinear prediction of river flows. *Journal of Hydrology*, (248):109–122.

Pulido-Calvo, I. and Portela, M. (2007). Application of neural approaches to one-step daily flow forecasting in portuguese watersheds. *Journal of Hydrology*, (332):1–15.

Ramadan, H., Beighley, R., and Ramamurthy, A. (2012). Modelling streamflow trends for a watershed with limited data: case of the litani basin, lebanon. *Hydrological Sciences Journal*, (57):1516–1529.

Rushworth, A., Bowman, A., Brewer, M., and Langan, S. (2013). Distributed lag models for hydrological data. *Biometrics*, (69):537–544.

Shanmuganathan, S. and Samarasinghe, S. (2016). Artificial neural network modelling. *Springer*, (628).

Sudheer, K., Gosain, A., and Ramasastri, K. (2002). A data-driven algorithm for constructing artificial neural network rainfall-runoff models. *Hydrol. Process*, (16):1325–1330.

Velasquez, J. and Palade, V. (2013). Advanced techniques in web intelligence: Web user browsing behaviour and preference analysis. *Springer*.

Verner, D., Lee, D., and Ashwill, M. (2013). Increasing resilience to climate change in the agricultural sector ofthe middle east: The cases of jordan and lebanon.

Vos, N. D. and Rientjes, T. (2005). Constraints of artificial neural networks for rainfall-runoff modelling: trade-offs in hydrological state representation and model evaluation. *Hydrology and Earth System Sciences*, (9):111–126.

Wu, C., Chau, K., and Fan, C. (2012). Prediction of rainfall time series using modular artificial neural networks coupled with data-preprocessing techniques. *Journal of Hydrology*, (389):146–167.

Wu, C., Chau, K., and Y.S, L. (2009). Methods to improve neural network performance in daily flows prediction. *Journal of Hydrology*, (69):537–544.