Hindawi Publishing Corporation
Advances in Artificial Intelligence
Volume 2014, Article ID 717803, 11 pages
http://dx.doi.org/10.1155/2014/717803

Research Article

Hybrid Wavelet-Postfix-GP Model for Rainfall Prediction of Anand Region of India

Vipul K. Dabhi[1] and Sanjay Chaudhary[2]

[1] *Information Technology Department, Dharmsinh Desai University, Nadiad 387001, India*
[2] *IICT, Ahmedabad University, Ahmedabad 380009, India*

Correspondence should be addressed to Vipul K. Dabhi; vipul.k.dabhi@gmail.com

Received 11 January 2014; Accepted 15 May 2014; Published 2 June 2014

Academic Editor: Djamel Bouchaffra

An accurate prediction of rainfall is crucial for national economy and management of water resources. The variability of rainfall in both time and space makes the rainfall prediction a challenging task. The present work investigates the applicability of a hybrid wavelet-postfix-GP model for daily rainfall prediction of Anand region using meteorological variables. The wavelet analysis is used as a data preprocessing technique to remove the stochastic (noise) component from the original time series of each meteorological variable. The Postfix-GP, a GP variant, and ANN are then employed to develop models for rainfall using newly generated subseries of meteorological variables. The developed models are then used for rainfall prediction. The out-of-sample prediction performance of Postfix-GP and ANN models is compared using statistical measures. The results are comparable and suggest that Postfix-GP could be explored as an alternative tool for rainfall prediction.

1. Introduction

An accurate prediction of rainfall is crucial for agriculture based Indian economy. Moreover, it also helps in the prevention of flood, the management of water resources, and generating recommendations related to crop for farmers [1]. The variability of rainfall in both time and space makes the rainfall prediction a challenging task. Moreover, the meteorological parameters needed for the rainfall prediction are complex and nonlinear in nature. Practitioners have applied numerical [2, 3] and statistical [4, 5] models for the rainfall prediction. The Numerical Weather Prediction (NWP) models are deterministic models and approximate complex physical processes for weather prediction. However, the models are not useful for prediction at smaller scale due to inherent limitations of these models to initial conditions and model parameterization. Practitioners have also used autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) techniques for developing a model for the rainfall [6]. However, these approaches were developed based on the assumption of stationarity of the given time series and the independence of the residuals. Moreover, these approaches lack the ability to identify nonlinear patterns and irregularity in the time series.

Hence, in recent years, use of different machine learning techniques for modeling and prediction of rainfall has received much attention of practitioners [7]. Hung et al. [8] developed a neural network for 1 to 3 hours ahead forecast of rainfall for Bangkok. They used meteorological parameters (air pressure, relative humidity, wet bulb temperature, and cloudiness) and the rainfall registered at the neighbor stations as input in the neural network. They used the hydrometeorological data that covers both rainy and nonrainy periods for training the neural network. Moustris et al. [9] used a neural network approach for forecasting the monthly minimum, maximum, mean, and cumulative precipitation for the four meteorological stations of Greece. They noted that ANN was not able to predict the peaks in all cases. They suggested increasing the size of training dataset to overcome this problem [9]. Lin and Chen [10] used a neural network for forecasting typhoon rainfall. ANN is developed which takes typhoon characteristics (direction of typhoon movement,

latitude and longitude of the typhoon centre, the maximum wind speed, and atmospheric pressure of the centre) and the spatial rainfall information of nearby rain gauge as inputs and gives 1 hour ahead forecast of typhoon rainfall. Practitioners applied ANN [11, 12] and approach based on chaos theory [13] for numerical model error prediction in hydrology. ANN and concepts of chaos theory are used to adjust the value of outputs produced by the numerical model.

The hydrometeorological time series can be considered as a composition of stochastic (noise or fluctuations) and structured components. The stochastic component obscures the modeling of time series. The structured component can be extracted by removing the stochastic component from a time series. Then, a deterministic model can be developed for the structured component of a time series. Practitioners have used following data preprocessing techniques for cleaning hydrological time series: wavelet analysis (WA), principal component analysis (PCA), and singular spectrum analysis (SSA). However, in recent years, the wavelet analysis has become an effective tool for analyzing nonstationary time series.

Partal and Cigizoglu [14] proposed a wavelet-ANN approach for predicting the daily precipitation of 12 meteorological stations of Turkey. They used meteorological data for precipitation prediction. Nasseri et al. [15] applied a combination of back propagation algorithm and genetic algorithm (GA) for rainfall forecasting in western suburbs of Sydney. They used GA to train and optimize feed-forward neural network. They concluded that a combined approach outperformed an approach that uses a neural network alone. A modular artificial neural network (MANN) [16] is combined with three data preprocessing techniques: moving average (MA), PCA, and SSA for prediction of two monthly and two daily precipitation series.

The symbolic regression technique can be used for developing a model (mathematical model in a symbolic form) that can explain the relationship between rainfall and meteorological parameters. The advantage of the symbolic regression technique over traditional regression techniques is that it searches for both the structure and the appropriate numeric coefficients of the model. Symbolic regression can be performed by means of genetic programming (GP) [17]. Moreover, GP approach is preferred over other approaches for symbolic regression because the approach produces an explicit mathematical expression as a solution (model) [18]. The produced model can provide an insight into the process which gives rise to the data [19]. Moreover, the interpretation of the produced model allows us to combine evolved knowledge with already existing knowledge [20, 21]. An exhaustive survey on open issues and approaches used by practitioners to deal with these issues in field of symbolic regression through GP is presented in [22]. Kisi and Shiri [23] suggested that use of GP is preferred in following situations: (i) the relationship between the relevant variables are poorly understood, (ii) determination of optimal solution is difficult, (iii) an approximate solution is acceptable, and (iv) there is a large amount of training data to be modeled.

Considering the mentioned advantages, many practitioners applied GP for rainfall prediction. GP is used in [24] to develop a model that can explain the cause and effect relationship between rainfall and runoff processes at a catchment in Singapore. Babovic and Keijzer [25] employed GP for developing rainfall-runoff models using the hydrometeorological data and the available domain knowledge. Khu et al. [26] applied GP for real-time runoff forecasting at the Orgeval catchment in France. The GP is used as an error updating strategy to accompany a rainfall-runoff model. The GP produced results are compared with those obtained using autoregression and Kalman filter. The results suggest that GP outperforms other strategies for real-time flow forecasting [26].

The objective of the present work is to explore the applicability of a hybrid wavelet-postfix-GP model for prediction of daily rainfall of Anand station of Gujarat, India, using meteorological variables. The wavelet analysis is used to decompose the original series of each meteorological variable into various discrete wavelet (DW) subseries. The decomposition is useful to identify the DW subseries which have high correlation with the original rainfall series. The effective DW subseries for each meteorological variable are added to generate a final subseries. The objective behind addition of DW subseries is to increase the correlation between the final subseries and the original rainfall series. The Postfix-GP, a GP variant, is then used to develop a model that can explain the relationship between the final subseries of meteorological variables and the original rainfall series. The Postfix-GP uses linear individual representation and stack based evaluation. This helps Postfix-GP to minimize both the memory requirement and evaluation time compared to conventional tree based representation. The evolved Postfix-GP models are then used for out-of-sample predictions. The performance of the evolved Postfix-GP model is measured using mean absolute error (MAE), mean squared error (MSE), and correlation coefficient (CC).

The rest of the paper is organized as follows. The next section presents homogeneity analysis of collected meteorological data. Section 3 presents the proposed hybrid wavelet-Postfix-GP model for rainfall prediction. Section 4 presents experimental settings. The evolved Postfix-GP model that represents rainfall as function of meteorological subseries is presented in Section 5. The section also presents comparison of the out-of-sample predictive performance of wavelet-Postfix-GP model and wavelet-ANN model. Conclusions are presented in Section 6.

2. Data and Homogeneity Analysis

Meteorological time series of Anand region for 12 years (from 1991 to 2002) were collected from Anand Agriculture University, Anand, Gujarat, India. We have used 10 years of data for training and 2 years of data for predictions. We have collected time series data for the following meteorological variables: (i) minimum temperature—T_{Min}, (ii) maximum temperature—T_{Max}, (iii) mean temperature—T_{Mean}, (iv) relative humidity—RH, (v) evaporation—EP, (vi) 1-day previous rainfall—RF_1, and (vii) 2-day previous rainfall—RF_2. The objective is to evolve a model (as a function of mentioned meteorological

Table 1: Result of different homogeneity tests.

Test	T_{Min}	T_{Max}	T_{Mean}	RH	EP	RF
Pettitt	0.204	0.241	0.238	0.354	0.481	0.093
SNHT	0.372	0.466	0.500	0.528	0.225	0.151
BR	0.287	0.438	0.395	0.508	0.455	0.115
VNR	0.495	0.146	0.595	0.004	0.304	0.127

variables) for daily rainfall using a hybrid wavelet-Postfix-GP approach.

Homogeneity tests are applied to detect the variability of the meteorological data. Several factors can affect the quality of meteorological data. The main sources of inhomogeneity are station relocation, changes in measurement techniques and observational procedures, and changes in instruments. Homogeneity test are useful to detect the break (shift in the mean) in the given time series. As many meteorological parameters are highly variable in time and space, most of the homogeneity tests are designed for monthly or yearly data and not for daily data. We applied the following four homogeneity tests to meteorological time series [27]: standard normal homogeneity test (SNHT), Buishand range (BR) test, Pettitt test, and von Neumann ratio (VNR) test. The null hypothesis for all the mentioned tests is the annual values X_i of the test variable X which are independent and identically distributed and series can be viewed as homogeneous [27]. The alternative hypothesis for BR test, SNHT, and Pettit test presumes that there is a break in the mean of the series and the series can be viewed as inhomogeneous. The reason for applying more than one test to check homogeneity is that these tests have different sensitivity in detecting a break. For example, the SNHT test is sensitive in detecting a break near the starting and the end of the series whereas Pettit and Buishand tests are sensitive in detecting break in the middle of the series.

The significance level is set to 1% for all tests. If the obtained P value is lower than the selected significance level, then we reject the null hypothesis. Table 1 presents the results of homogeneity tests for the annual mean values of meteorological parameters. As all the tests give values greater than 0.01, the null hypothesis cannot be rejected (time series is homogeneous). For the relative humidity, VNR test gives a P value of 0.004, less than 0.01. However, as the P values associated to the other tests are greater than 0.01, we accept the null hypothesis.

3. A Hybrid Wavelet Postfix-GP Model

The architecture that integrates discrete wavelet transform (DWT) and Postfix-GP for modeling and prediction of rainfall time series is presented in Figure 1. The architecture comprises the following main steps: (i) apply DWT on meteorological data and generate DW subseries at every level, (ii) calculate the correlation coefficient between the generated DW subseries and the original rainfall series, (iii) identify the significant DW subseries, (iv) add significant DW subseries to generate a new subseries for each meteorological variable, (v) normalize the values of newly generated subseries in the range (0, 1), (vi) divide the normalized dataset into (a) training and (b) test dataset, (vii) evolve the model for training data using Postfix-GP, and (viii) apply the evolved model for out-of-sample rainfall prediction. The objective is to evolve a model (as a function of mentioned meteorological variables) for rainfall using a hybrid wavelet-Postfix-GP approach. In the following subsections, we discuss the wavelet transform and Postfix-GP in brief.

3.1. Wavelet Transform. The properties of irregularity in shape and compactness make wavelets an ideal tool for analysis of nonstationary signals. Fourier analysis decomposes a signal into sine and cosine waves of various frequencies whereas wavelet analysis decomposes a signal into shifted and scaled versions of the mother wavelet. The shifting (delaying) of the mother wavelet provides local information of the signal in time domain whereas scaling (stretching or compressing) of the mother wavelet provides local information of the signal in frequency domain [28]. The scaling and shifting operations applied to mother wavelet are used to calculate wavelet coefficients that provide correlation between the wavelet and local portion of the signal. From the calculated wavelet coefficients, we can extract two types of components: approximate coefficients and detail coefficients. The approximate coefficients represent high scale, low frequency component of the original signal whereas detail coefficients represent low scale, high frequency component.

Continuous wavelet transform (CWT) operates at every scale from that of the original signal up to some maximum scale. This distinguishes CWT from DWT (which operates at dyadic scales only). CWT is also continuous in terms of shifting: during computation, the analyzing wavelet is shifted smoothly over the full domain of signal. The results of the CWT are wavelet coefficients, which are a function of scale and position. Multiplying each coefficient by the appropriately scaled and shifted wavelet gives the constituent wavelets of the original signal.

The computation of wavelet coefficients at every scale requires large computational time. To reduce the time, it is preferred to calculate wavelet coefficients for selected subset of scales and positions. If the scales and positions are selected based on power of two (dyadic scales and positions), then the analysis will be efficient and just as accurate, named discrete wavelet transform (DWT) [29]. The process of decomposition can be iterated, with successive approximations being decomposed in turn (discarding detail coefficients), so that original signal is broken down into many lower-resolution components. This process is referred

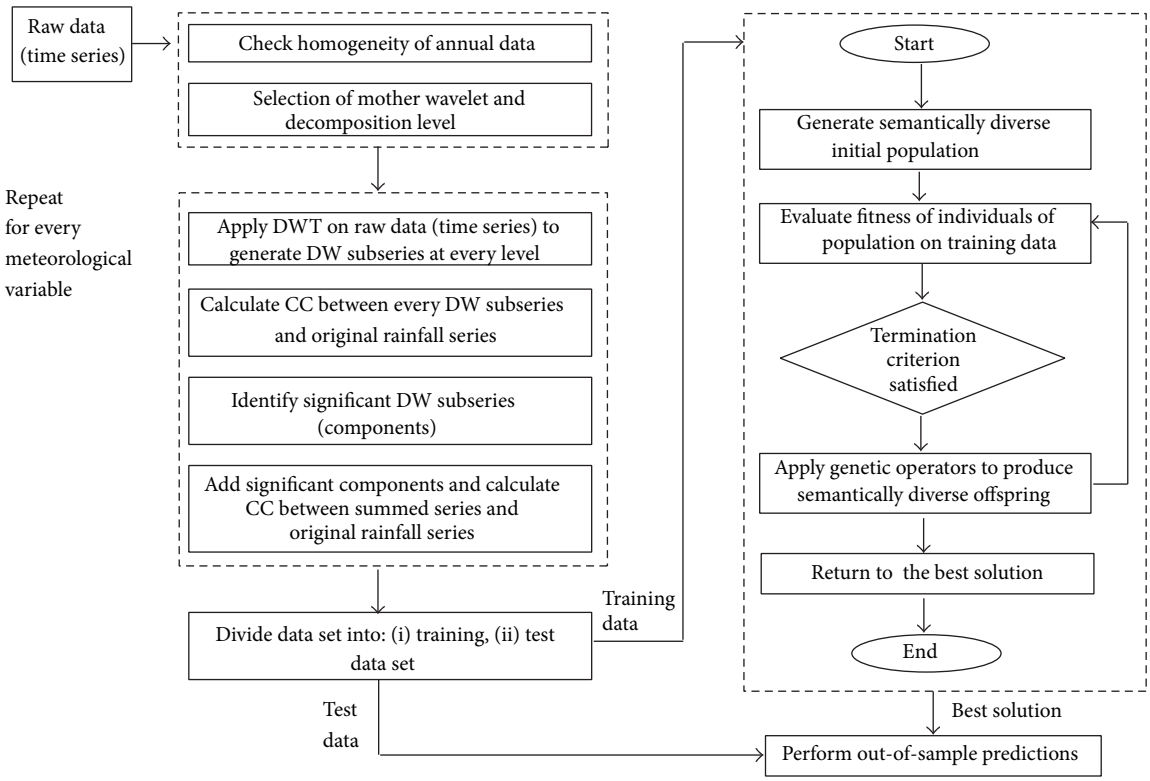

Figure 1: Hybrid Wavelet-Postfix-GP architecture for rainfall modeling and prediction.

as multiresolution analysis. We have selected Db4 (length-4 Daubechies) [28] wavelet as mother wavelet because this is one of the commonly used wavelets for separating fluctuations from the given time series. The smoothness of different wavelets depends on the number of vanishing moments [29]. Db4 wavelet has four vanishing moments, a smallest length wavelet with smoothness property. We have set maximum resolution level to value 10 for decomposition of every meteorological time series.

The forward discrete wavelet transform is employed to decompose original time series of every meteorological variable at different scale (maximum level $n = 10$). The wavelet transform produces high-pass (detail) coefficients at every level and one low-pass (approximation) coefficient. The inverse discrete wavelet transform is applied on the produced coefficients at every level to generate DW subseries, which are of the same length as the original series. Thus, we obtain $n + 1$ (n detail and one approximate) DW subseries for the original time series of every meteorological variable. The correlation coefficient between the generated DW subseries at different level and the original rainfall series is calculated. The number of DW subseries which have high correlation with the original rainfall series is identified and summed up to generate a new (final) subseries for that meteorological variable. The objective behind addition of DW subseries having high correlation with the original rainfall series is to reduce the number of variables (dimensions or inputs) and to increase the correlation between newly generated subseries and the original rainfall series. This process is repeated for every meteorological time series.

3.2. Postfix Genetic Programming. Postfix-GP [30], a GP variant, adopts postfix notation for individual representation. Individuals represented in form of postfix strings can be easily evaluated using stack. Moreover, the individual representation without pointers assists Postfix-GP to minimize the fitness evaluation time and required memory to store the individual. Each Postfix-GP individual contains the following three attributes: MinLength, MaxLength, and ValidLength. The MinLength and MaxLength attributes define the range of syntactically valid Postfix-GP individuals. The ValidLength attribute refers to the index of the last element of an individual forming a valid postfix expression. An individual with ValidLength greater than MinLength and less than MaxLength is considered as valid [30].

Postfix-GP employs idea of Stack count, introduced by Keith and Martin [31], to find out the ValidLength of an individual. The Stack count of an element is calculated as the number of arguments pushed on the stack minus the number of arguments popped off the stack by the element. For example, the stack count value for an operand is 1, whereas it is 0 for unary operator. Furthermore, the total sum of stack count values must be 1 at the ValidLength position of an individual. A Postfix-GP individual with its syntactically valid portion and corresponding tree representation is depicted in Figure 2. It should be noted that the transformation of syntactically valid portion to a tree representation is not required for fitness calculation; it is shown for better comprehension of the reader.

Table 2 presents difference between an individual representation scheme of Postfix-GP and other evolutionary

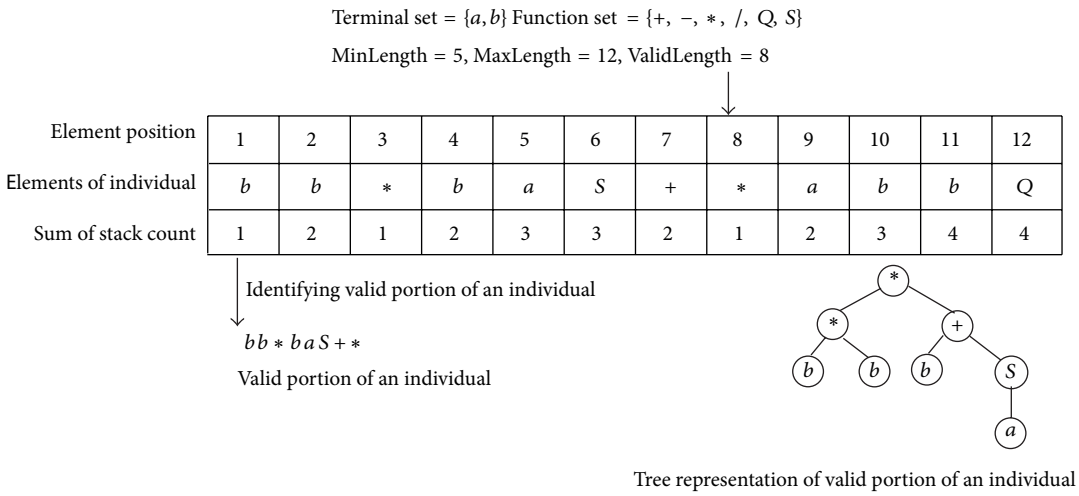

Figure 2: Postfix-GP individual representation.

Table 2: Difference between individual representation schemes of Postfix-GP and other evolutionary approaches.

Evolutionary algorithms	Genotype	Phenotype	Mapping
GA	Linear string of fixed length	Decoding genotype	Problem dependent
GP	Parse tree	Parse tree	Traversing tree
GEP	Linear string of fixed length	Parse tree	Karva notation
Postfix-GP	Linear string of fixed length	Linear string	Postfix notation

approaches, GA [32], GP [17], and gene expression programming (GEP) [33]. The standard GA represents an individual using a fixed length binary string. The individual representation scheme of GA does not permit the model structure to vary during evolutionary process. Standard GP employs a variable length, tree structure for an individual representation. The representation scheme of GP is more general and flexible than GA since it allows model structure to vary during evolution. GEP is a GP variant that represents an individual using multiple genes, where each gene represents a small subexpression. Each GEP gene is composed of two different domains—a head and a tail. GEP applies Karva notation to transform a linear representation to tree representation. Grammatical evolution (GE) [34] employs a string of integers to represent an individual (genotype). The string of integers is used to determine the sequence of production rules in context-free grammar. By following these sequences, an individual can be converted into an expression tree (phenotype). Adaptive logic programming (ALP) [35] also uses a string of integers for individual representation. However, the string of integers is used to select clauses in a logic program instead of production rules of a context-free grammar.

Postfix-GP produces initial population of individuals at random. The ValidLength of these individuals will have value in between MinLength and MaxLength. Postfix-GP employs semantic aware subtree crossover [36] to improve population diversity among individuals. The operator checks semantic equivalence [37, 38] of two subtrees, to be swapped, while performing crossover operation. Moreover, the operator selects behaviorally different parents for generating offspring, which is useful to minimize the "no change to fitness" events [39]. Crossover of two dissimilar parents is likely to produce a change in offspring (solution) quality.

Postfix-GP extracts all subtrees of an individual having ValidLength greater than MinLength and treat the extracted subtrees as separate solutions during the evolutionary process. Postfix-GP employs one-point mutation operator, where the chosen element of an individual is interchanged with a different element of the same arity. An archive is used to store "best-so-far" found solutions, useful to exploit good solutions over a number of generations [40]. Postfix-GP uses MAE as a standardized fitness measure. However, MAE measure is not range bound. Therefore, we have normalized the fitness value of an individual between 0 and 1 using (1). The normalized fitness is referred to as an adjusted fitness of an individual and is useful to differentiate between individuals having close standardized fitness values, which may happen in generations near the end of Postfix-GP run. Postfix-GP is developed using.NET [41] framework on Windows XP operating system. Zedgraph [42], an open source graph library, is used for plotting charts.

$$\text{Adjusted Fitness} = \frac{1}{(1 + \text{Standardized Fitness})}. \quad (1)$$

4. Experimental Settings

We set a range of solution search space by specifying a minimum and maximum number of elements (nodes) that

Table 3: Correlation coefficient between each subseries and the original rainfall series.

Subseries	T_{Min}	T_{Max}	T_{Mean}	RH	EP	RF_1	RF_2
DW_1	−0.079	−0.029	−0.076	0.033	−0.127	−0.304	−0.013
DW_2	−0.068	−0.093	−0.103	0.110	−0.124	0.130	−0.249
DW_3	−0.060	−0.098	−0.097	0.101	−0.156	0.355	0.187
DW_4	−0.033	−0.098	−0.087	0.114	−0.177	0.302	0.258
DW_5	−0.027	−0.115	−0.094	0.132	−0.165	0.273	0.259
DW_6	−0.043	−0.053	−0.056	0.052	−0.065	0.138	0.135
DW_7	−0.030	−0.145	−0.102	0.160	−0.153	0.211	0.210
DW_8	0.230	0.117	0.211	0.225	0.025	0.251	0.250
DW_9	0.013	−0.045	−0.013	0.072	−0.044	0.081	0.081
DW_{10}	−0.016	−0.018	−0.017	0.019	−0.026	0.019	0.019
Approximate	0.018	0.004	0.014	0.034	0.008	0.044	0.044
Summed	0.230	−0.236	0.211	0.356	−0.347	0.613	0.519

a solution can have during an evolutionary run. The MinLength and MaxLength parameters of Postfix-GP are used to attain this task. The MinLength and MaxLength parameters of Postfix-GP are set to values 10 and 40. The function set includes both arithmetic and trigonometric operators: {+, −, /, ∗, S, C, E, L, K, Q}, where S, C, E, L, and Q represent sine, cosine, exponential, logarithmic, and square root functions. The terminal set includes final subseries for T_{Min}, T_{Max}, T_{Mean}, RH, EP, RF_1, RF_2} and a list of constants in range [−10,. . .,10]. The population size and the number of generations are set to values 500 and 100. The crossover and mutation rates are set to 0.9 and 0.1. The Postfix-GP uses archive based roulette wheel selection [40]. Crossover operation is performed between parents chosen from an archive and current population

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\left(y_i - \widehat{y}_i\right)\right|.$$

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \widehat{y}_i\right)^2, \tag{2}$$

$$r = \frac{\text{cov}\left(Y, \widehat{Y}\right)}{\sigma_Y \sigma_{\widehat{Y}}},$$

We have shown the evolved Postfix-GP solutions with their mean absolute error (MAE), mean squared error (MSE), and correlation coefficient (CC). Equations in (2) are used to calculate these statistical measures, where y_i, $\widehat{y}_i$, Y, $\widehat{Y}$, and $\text{cov}(Y, \widehat{Y})$ stand for given ith observation, corresponding ith observation calculated by the evolved model, actual observed series, estimated series by the evolved model, and covariance between the observed and the estimated series. Value of r close to zero indicates no correlation between predicted and observed values. We have used MSE to measure closeness between the given set of points and those generated by the evolved solution.

5. Results

5.1. Identification of Important Components. The discrete wavelet transform (DWT) is applied on daily time series of every meteorological variable to decompose the series into several DW subseries [14]. Each time series is decomposed up to 10 resolution levels. We have selected the resolution level of 10 because we required addressing both annual and seasonal factors. The levels DW_2, DW_3, DW_4, DW_5, DW_6, DW_7, DW_8, DW_9, and DW_{10} correspond to the temporal scale of 4, 8, 16, 32, 64, 128, 256, 512, and 1024 days. The subseries, DW_8, correspond to a temporal scale of 256 days (approximately one year). The correlation coefficient between the generated DW subseries at different levels and the original rainfall series is calculated. Table 3 presents the values of the correlation coefficient between each DW subseries and the original rainfall series. These correlation values are used to determine the effective DW components (subseries) for rainfall prediction.

According to Table 3, for minimum and mean temperature, the correlation between DW_8 subseries and the original rainfall series has high value. This reveals the fact that annual component of mean and minimum temperature subseries influences the amount of rainfall. The DW_7 subseries of maximum temperature has high correlation compared to other DW subseries. Moreover, the DW_2, DW_3, DW_4, and DW_5 subseries of maximum temperature have little high correlation compared to other DW subseries. Similarly, the DW_1, DW_2, DW_3, DW_4, DW_5, and DW_7 subseries of evaporation have high correlation compared to other DW subseries. This suggests that shorter time period components of maximum temperature and evaporation have correlation with the original rainfall series. The high correlation value is noticed for DW_8 subseries of relative humidity. Moreover, the correlation between the original rainfall and DW_2, DW_3, DW_4, DW_5, and DW_7 subseries of relative humidity is little high compared to correlation between the rainfall and remaining DW subseries of relative humidity. This suggests that both annual and shorter time period components of relative humidity have correlation with the rainfall.

Table 4: Selected DW components for generating final subseries.

T_{Min}	T_{Max}	T_{Mean}	RH	EP	RF_1	RF_2
DW_8	$DW_2 + DW_3 + DW_4 + DW_5 + DW_7$	DW_8	$DW_2 + DW_3 + DW_4 + DW_5 + DW_7 + DW_8$	$DW_1 + DW_2 + DW_3 + DW_4 + DW_5 + DW_7$	$DW_2 + DW_3 + DW_4 + DW_5 + DW_6 + DW_7 + DW_8$	$DW_3 + DW_4 + DW_5 + DW_6 + DW_7 + DW_8$

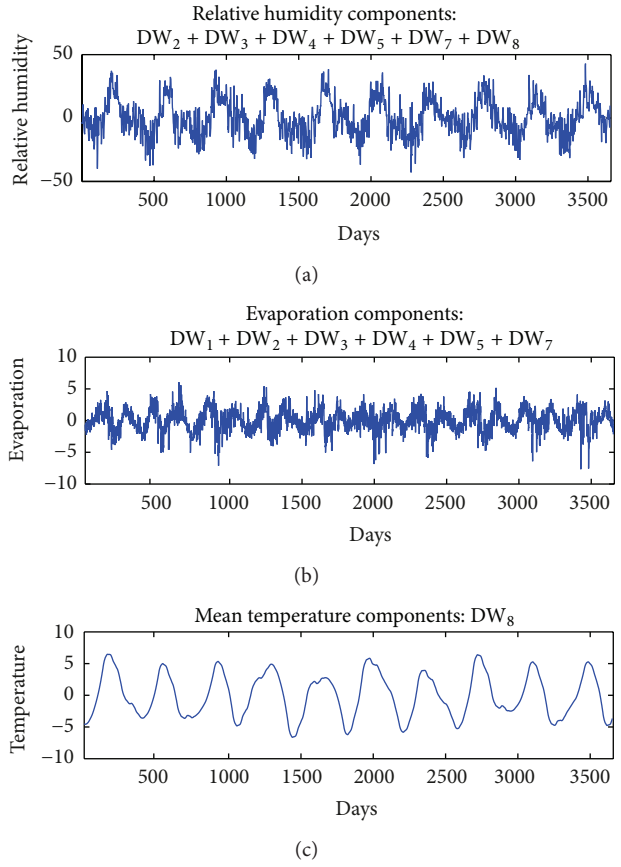

Figure 3: Summed subseries: (a) relative humidity, (b) evaporation, and (c) mean temperature.

The high correlation between both shorter and yearly components of relative humidity and rainfall is observed because of direct relation between these two variables. The humidity and rainfall data of the examined region show similar meteorological characteristics. The DW_3 subseries of one-day previous rainfall show high correlation with the original rainfall series. The DW_2, DW_4, DW_5, DW_6, DW_7, and DW_8 subseries also show little high correlation compared to other DW subseries. The number of DW subseries which have high correlation with the original rainfall series is identified and summed up to generate a new (final) subseries for that meteorological variable. The selected DW components for different meteorological variables are presented in Table 4. However, the selection of number of DW components (subseries) depends on the user. Partal and Cigizoglu [14] suggested that applying a threshold on correlation value would be helpful to determine the number of DW components (subseries).

Usage of each selected DW subseries as input increases the dimension of the solution search space. Moreover, it also increases the complexity of the final model. Therefore, the selected DW subseries, having high correlation with the original rainfall series, are added together to form a new (final) subseries. The final subseries are generated for every meteorological variable. The addition of selected DW components (subseries) is helpful to improve the correlation between the final subseries and the original rainfall series. For example, the DW_8 subseries of relative humidity have correlation value of 0.225. However, for the added ($DW_2 + DW_3 + DW_4 + DW_5 + DW_7 + DW_8$) subseries, the correlation value increases to 0.355.

The added subseries of the relative humidity, evaporation, and mean temperature are shown in Figure 3. The summed series of the maximum temperature, minimum temperature, and previous day rainfall are presented in Figure 4. We normalized the data of added subseries in the range (0, 1).

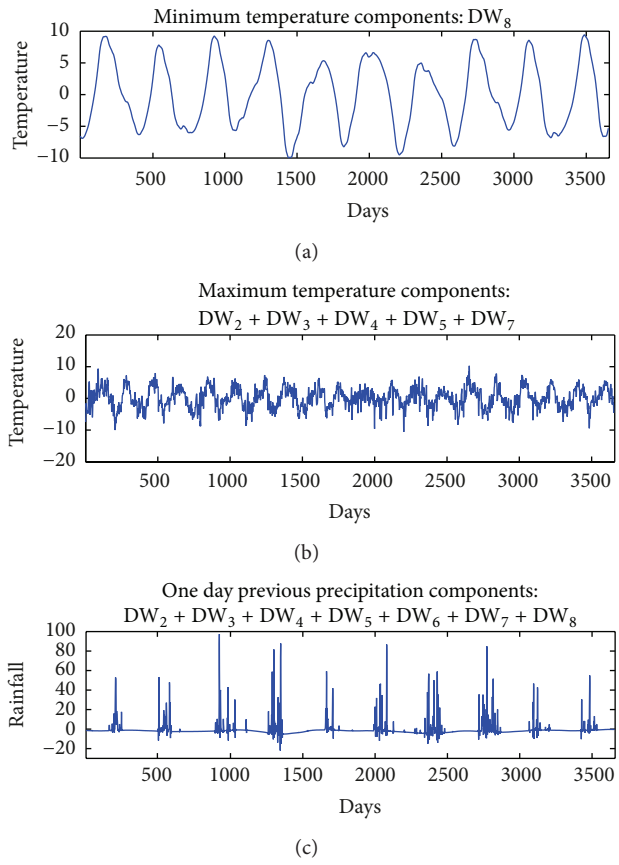

Figure 4: Summed subseries: (a) minimum temperature, (b) maximum temperature, and (c) 1-day previous rainfall.

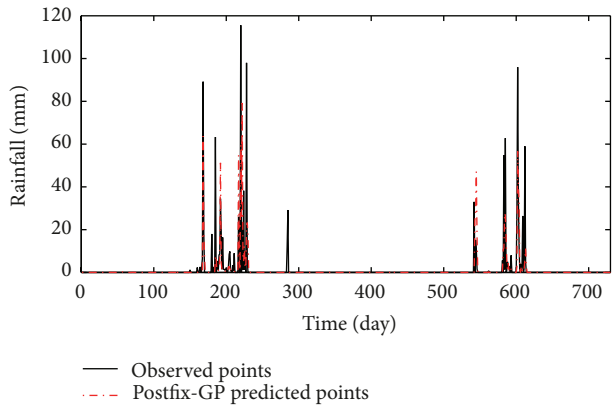

Figure 5: Daily rainfall prediction by Postfix-GP Model for the test period (2 years).

The data normalization assures that large value variables do not overwhelm small value variables.

5.2. Evolved Postfix-GP Solutions. The evolved Postfix-GP solution (for terminal set $\{T_{\text{Min}}, T_{\text{Max}}, T_{\text{Mean}}, \text{EP}, \text{RH}, \text{RF}_1\}$) is shown in (3). It is noted that the solution contains significant meteorological variables (evaporation, minimum temperature, relative humidity, and maximum temperature) which are useful for rainfall prediction. We got MSE = 48.2860, adjusted fitness = 0.3926, and $r = 0.7469$ for the training data

$$e^{(\cos(\cos(-10.6\,\text{EP}\,T_i))-(X_i/(\sin(\text{RF}_1\,\text{RH})*\text{RH}))+3.7292)}, \quad (3)$$

where $X_i = \text{EP} * e^{\text{Log}_{10}(4.3\,\text{EP}+0.4941)}/(T_x + \text{RH})$.

The observed and Postfix-GP model predicted values for the testing period are presented in Figure 5. The model has predicted the general behavior of the observed rainfall data. It has accurately predicted the summer days and estimated zero rainfall for these days. However, the model performs satisfactory for estimating maximum rainfall values.

5.3. Comparison of Result Obtained by Wavelet-Postfix-GP and Wavelet-ANN Models. We have compared the performance of wavelet-Postfix-GP and wavelet-ANN models for the daily rainfall prediction. The selection of architecture (number of layers and number of nodes in each layer) and the training algorithm is important design parameters of ANN. There is

Table 5: Statistical measures for the best wavelet-Postfix-GP and wavelet-ANN models for different input combinations for test period.

Model inputs	Wavelet-Postfix-GP			Wavelet-ANN			
	MSE	Adj_{Fit}	R	ANN structure	MSE	Adj_{Fit}	R
T_{Min}, T_{Max}, T_{Mean}, EP, RH, RF_1, RF_2	49.2494	0.4002	0.6661	7, 5, 1	52.6614	0.2804	0.6919
T_{Min}, T_{Max}, T_{Mean}, EP, RH, RF_1	47.5766	0.4017	0.6794	6, 6, 1	50.4010	0.3545	0.7641

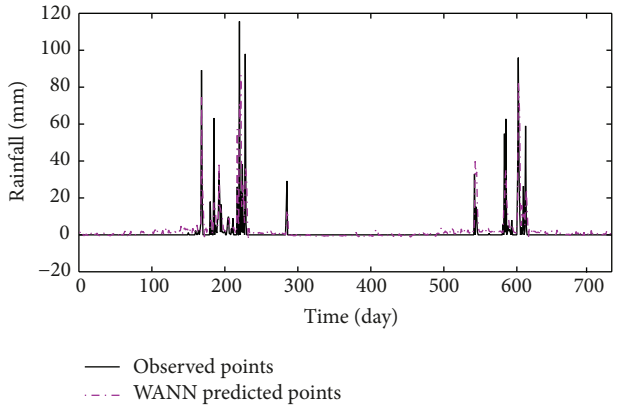

Figure 6: Daily rainfall prediction by ANN model for the test period (2 years).

no rule available to find out the number of hidden layers and appropriate number of nodes for each hidden layer. However, practitioners [43] found that ANN with one hidden layer is complex enough to model the nonlinear properties of the hydrologic processes. We investigated different architecture with different activation functions and number of neurons for the hidden layer and selected the one which gives an optimal result. A multilayer feed-forward ANN that comprises seven nodes in the input layer, one hidden layer with five nodes, and one node in the output layer is selected for the rainfall prediction. Moreover, the input and hidden layers have an additional bias neuron. The ANN is trained with back propagation (BP) learning algorithm. We set log-sigmoid transfer function between input and hidden layers and pure linear transfer function between hidden and output layers.

Figure 6 presents the observed and wavelet-ANN model predicted rainfall values. The model has predicted the general behavior of the observed rainfall data. It is observed that wavelet-ANN model has produced negative prediction for some days having zero rainfall, which are not practically possible. Partal and Cigizoglu [14] found the similar behavior while applying ANN for predicting daily precipitation. They noted that the negative prediction problem occurred due to extrapolation ability of feed-forward back-propagation mechanism. Moreover, similar to Postfix-GP, the ANN failed in accurately predicting the peak values of rainfall during the test period.

Table 5 presents the statistical results for both wavelet-Postfix-GP and wavelet-ANN hybrid models for different input combinations for the test period. The Postfix-GP model has produced slightly better results than ANN from the MSE and fitness viewpoint. The ANN model slightly outperforms Postfix-GP in terms of correlation coefficient. The results suggest that Postfix-GP produces comparable results and can be an alternative to ANN approach.

6. Conclusion

We applied four homogeneity tests (SNHT, BR, Pettit, and VNR) to detect the variability of the meteorological (minimum temperature, mean temperature, maximum temperature, evaporation, and relative humidity) variables recorded at Anand station. The results of homogeneity tests suggest that the time series of the meteorological variables are homogeneous. The series of every meteorological variable is decomposed up to 10 resolution level using discrete wavelet transform. The correlation coefficient between the generated DW subseries at different levels and the original rainfall series is calculated. The number of DW subseries which have correlation with the original rainfall series is identified and summed up to generate a new subseries for every meteorological variable. It is observed that both annual and shorter time period components of relative humidity have correlation with rainfall. The annual components of minimum and mean temperature show correlation with rainfall series. The newly generated meteorological subseries are regarded as inputs and the rainfall series as output.

The Postfix-GP is then employed to develop a model that can explain relationship between the inputs and the output. The developed model is then used for the rainfall prediction. The prediction performance of the evolved model was good, giving low values for MAE and MSE and high value for the correlation coefficient, suggesting that Postfix-GP is able to evolve an accurate and reliable model. The advantage of Postfix-GP over ANN is that it gives an explicit mathematical nonlinear equation that describes the relationship between inputs and output. The predictive performance of Postfix-GP and ANN models is compared. The results show that the Postfix-GP is a fair competitor of ANN approach. Moreover, the predictive performance of the evolved Postfix-GP model for the nonrainy (zero rainfall) periods and rainy (highest rainfall) periods is satisfactory as compared to ANN model, which produces negative prediction for some days of summer season. We conclude that wavelet-Postfix-GP approach obtained good quality solutions for the tested rainfall prediction series and could be explored as an alternative tool for predicting the hydrometeorological variables. Our future plan is to apply Postfix-GP for developing more accurate and reliable models using different combination of function and terminal sets.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors would like to acknowledge Krishi bhavan, Gandhinagar, and Anand Agricultural University, Anand, for their support in terms of meteorological data.

References

[1] V. Y. Jain, A. Sharma, S. Chaudhary, and V. K. Tyagi, "Spatial analysis for generating recommendations for agricultural crop production," in *Proceedings of the India Conference on Geospatial Technologies and Applications (ICGTA '12)*, 2012.

[2] L. R. Nayagam, R. Janardanan, and H. S. R. Mohan, "An empirical model for the seasonal prediction of southwest monsoon rainfall over Kerala, a meteorological subdivisionof India," *International Journal of Climatology*, vol. 28, no. 6, pp. 823–831, 2008.

[3] T. DelSole and J. Shukla, "Linear prediction of Indian monsoon rainfall," *Journal of Climate*, vol. 15, no. 24, pp. 3645–3658, 2002.

[4] A. R. Ganguly and R. L. Bras, "Distributed quantitative precipitation forecasting using information from radar and numerical weather prediction models," *Journal of Hydrometeorology*, vol. 4, no. 6, pp. 1168–1180, 2003.

[5] T. Diomede, S. Davolio, C. Marsigli et al., "Discharge prediction based on multi-model precipitation forecasts," *Meteorology and Atmospheric Physics*, vol. 101, no. 3-4, pp. 245–265, 2008.

[6] C. Chatfield, *The Analysis of Time Series: An Introduction*, CRC press, New York, NY, USA, 2003.

[7] V. Babovic, "Data mining in hydrology," *Hydrological Processes*, vol. 19, no. 7, pp. 1511–1515, 2005.

[8] N. Q. Hung, M. S. Babel, S. Weesakul, and N. K. Tripathi, "An artificial neural network model for rainfall forecasting in Bangkok, Thailand," *Hydrology and Earth System Sciences*, vol. 13, no. 8, pp. 1413–1425, 2009.

[9] K. P. Moustris, I. K. Larissi, P. T. Nastos, and A. G. Paliatsos, "Precipitation forecast using artificial neural networks in specific regions of Greece," *Water Resources Management*, vol. 25, no. 8, pp. 1979–1993, 2011.

[10] G.-F. Lin and L.-H. Chen, "Application of an artificial neural network to typhoon rainfall forecasting," *Hydrological Processes*, vol. 19, no. 9, pp. 1825–1837, 2005.

[11] V. Babovic, R. Caňizares, H. R. Jensen, and A. Klinting, "Neural networks as routine for error updating of numerical models," *Journal of Hydraulic Engineering*, vol. 127, no. 3, pp. 181–193, 2001.

[12] Y. Sun, V. Babovic, and E. S. Chan, "Multi-step-ahead model error prediction using time-delay neural networks combined with chaos theory," *Journal of Hydrology*, vol. 395, no. 1-2, pp. 109–116, 2010.

[13] V. Babovic, S. A. Sannasiraj, and E. S. Chan, "Error correction of a predictive ocean wave model using local model approximation," *Journal of Marine Systems*, vol. 53, no. 1–4, pp. 1–17, 2005.

[14] T. Partal and H. K. Cigizoglu, "Prediction of daily precipitation using wavelet-neural networks," *Hydrological Sciences Journal*, vol. 54, no. 2, pp. 234–246, 2009.

[15] M. Nasseri, K. Asghari, and M. J. Abedini, "Optimized scenario for rainfall forecasting using genetic algorithm coupled with artificial neural network," *Expert Systems with Applications*, vol. 35, no. 3, pp. 1415–1421, 2008.

[16] C. Wu, K. Chau, and C. Fan, "Prediction of rainfall time series using modular artificial neural networks coupled with data-preprocessing techniques," *Journal of Hydrology*, vol. 389, no. 1-2, pp. 146–167, 2010.

[17] J. R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, MIT Press, Cambridge, Mass, USA, 1992.

[18] V. Babovic, "Introducing knowledge into learning based on genetic programming," *Journal of Hydroinformatics*, vol. 11, no. 3-4, pp. 181–193, 2009.

[19] M. Keijzer and V. Babovic, "Declarative and preferential bias in GP-based scientific discovery," *Genetic Programming and Evolvable Machines*, vol. 3, no. 1, pp. 41–79, 2002.

[20] V. Babovic and M. B. Abbott, "The evolution of equations from hydraulic data Part I: theory," *Journal of Hydraulic Research*, vol. 35, no. 3, pp. 397–430, 1997.

[21] V. Babovic and M. B. Abbott, "Evolution of equations from hydraulic data. Part II: applications," *Journal of Hydraulic Research*, vol. 35, no. 3, pp. 411–430, 1997.

[22] V. K. Dabhi and S. Chaudhary, "Empirical modeling using genetic programming: a survey of issues and approaches," *Natural Computing*, 2014.

[23] O. Kisi and J. Shiri, "Precipitation forecasting using wavelet-genetic programming and wavelet-neuro-fuzzy conjunction models," *Water Resources Management*, vol. 25, no. 13, pp. 3135–3152, 2011.

[24] S.-Y. Liong, T. R. Gautam, T. K. Soon, V. Babovic, M. Keijzer, and N. Muttil, "Genetic Programming: a new paradigm in rainfall runoff modeling," *Journal of the American Water Resources Association*, vol. 38, no. 3, pp. 705–718, 2002.

[25] V. Babovic and M. Keijzer, "Rainfall-runoff modelling based on genetic programming," *Nordic Hydrology*, vol. 33, no. 5, pp. 331–346, 2002.

[26] S. T. Khu, S.-Y. Liong, V. Babovic, H. Madsen, and N. Muttil, "Genetic programming and its application in real-time runoff forecasting," *Journal of the American Water Resources Association*, vol. 37, no. 2, pp. 439–451, 2001.

[27] J. B. Wijngaard, A. M. G. Klein Tank, and G. P. Können, "Homogeneity of 20th century European daily temperature and precipitation series," *International Journal of Climatology*, vol. 23, no. 6, pp. 679–692, 2003.

[28] I. Daubechies, *Ten Lectures on Wavelets*, vol. 61, SIAM, Philadelphia, Pa, USA, 1992.

[29] S. Mallat, "A wavelet tour of signal processing," *A Wavelet Tour of Signal Processing*, 2009.

[30] V. K. Dabhi and S. K. Vij, "Empirical modeling using symbolic regression via postfix genetic programming," in *Proceedings of the International Conference on Image Information Processing (ICIIP '11)*, pp. 1–6, November 2011.

[31] M. J. Keith and M. C. Martin, "Genetic programming in c++: implementation issues," in *Advances in Genetic Programming*, pp. 285–310, 1994.

[32] J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*, MIT Press, Cambridge, Mass, USA, 1992.

[33] C. Ferreira, "Gene expression programming: a new adaptive algorithm for solving problems," *Complex Systems*, vol. 13, no. 2, pp. 87–129, 2001.

[34] M. O'Neil and C. Ryan, "Grammatical evolution," in *Grammatical Evolution*, pp. 33–47, Springer, New York, NY, USA, 2003.

[35] M. Keijzer, V. Babovic, C. Ryan, M. O'Neill, and M. Cattolico, "Adaptive logic programming," in *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '01)*, pp. 42–49, 2001.

[36] V. K. Dabhi and S. Chaudhary, "Semantic sub-tree crossover operator for postfix genetic programming," *Advances in Intelligent Systems and Computing*, vol. 201, no. 1, pp. 391–402, 2013.

[37] N. Q. Uy, N. X. Hoai, and M. O'Neill, "Semantic aware crossover for genetic programming: the case for real-valued function regression," in *Proceedings of the 12th European Conference on Genetic Programming (EuroGP '09)*, pp. 292–302, Springer, Berlin, Germany, 2009.

[38] N. Q. Uy, N. X. Hoai, M. O'Neill, R. I. McKay, and E. Galván-López, "Semantically-based crossover in genetic programming: application to real-valued symbolic regression," *Genetic Programming and Evolvable Machines*, vol. 12, no. 2, pp. 91–119, 2011.

[39] S. Gustafson, E. K. Burke, and N. Krasnogor, "On improving genetic programming for symbolic regression," in *Proceedings of the IEEE Congress on Evolutionary Computation (IEEE CEC '05)*, vol. 1, pp. 912–919, September 2005.

[40] M. Laumanns, L. Thiele, E. Zitzler, and K. Deb, "Archiving with guaranteed convergence and diversity in multi-objective optimization," in *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '02)*, pp. 439–447, Morgan Kaufmann, San Francisco, Calif, USA, 2002.

[41] Microsoft, Microsoft.net framework software development kit, 2007, http://msdn.microsoft.com/ .

[42] Zedgraph, 2008, http://sourceforge.net/projects/zedgraph/ .

[43] N. J. De Vos and T. Rientjes, "Constraints of artificial neural networks for rainfall-runoff modelling: trade-offs in hydrological state representation and model evaluation," *Hydrology and Earth System Sciences*, vol. 9, no. 1-2, pp. 111–126, 2005.