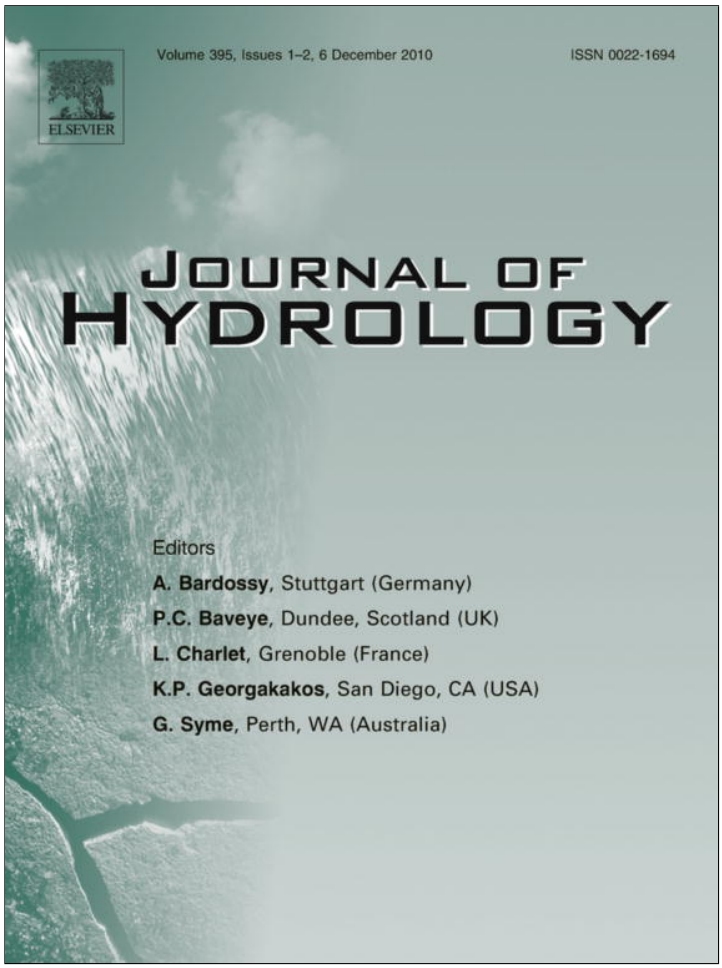

Journal of Hydrology 395 (2010) 23–38

Contents lists available at ScienceDirect

Journal of Hydrology

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

Streamflow prediction using multi-site rainfall obtained from hydroclimatic teleconnection

S.S. Kashid [a], Subimal Ghosh [a,*], Rajib Maity [b]

[a] Department of Civil Engineering, Indian Institute of Technology Bombay, Powai, Mumbai 400 076, India
[b] Department of Civil Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721 302, West Bengal, India

ARTICLE INFO

Article history:
Received 6 February 2010
Received in revised form 21 September 2010
Accepted 4 October 2010

This manuscript was handled by K. Georgakakos, Editor-in-Chief, with the assistance of Ercan Kahya, Associate Editor

Keywords:
El Niño Southern Oscillation (ENSO)
Equatorial Indian Ocean Oscillation (EQUINOO)
Outgoing Longwave Radiation (OLR)
Mahanadi River
Genetic Programming
Hydroclimatic teleconnection

SUMMARY

Simultaneous variations in weather and climate over widely separated regions are commonly known as "hydroclimatic teleconnections". Rainfall and runoff patterns, over continents, are found to be significantly teleconnected, with large-scale circulation patterns, through such hydroclimatic teleconnections. Though such teleconnections exist in nature, it is very difficult to model them, due to their inherent complexity.

Statistical techniques and Artificial Intelligence (AI) tools gain popularity in modeling hydroclimatic teleconnection, based on their ability, in capturing the complicated relationship between the predictors (e.g. sea surface temperatures) and predictand (e.g., rainfall). Genetic Programming is such an AI tool, which is capable of capturing nonlinear relationship, between predictor and predictand, due to its flexible functional structure. In the present study, gridded multi-site weekly rainfall is predicted from El Niño Southern Oscillation (ENSO) indices, Equatorial Indian Ocean Oscillation (EQUINOO) indices, Outgoing Longwave Radiation (OLR) and lag rainfall at grid points, over the catchment, using Genetic Programming.

The predicted rainfall is further used in a Genetic Programming model to predict streamflows. The model is applied for weekly forecasting of streamflow in Mahanadi River, India, and satisfactory performance is observed.

1. Introduction

Basin-scale streamflow prediction is an important step in water resources management for sustainable development. The variation of basin-scale streamflow is influenced by rainfall depth, its distribution pattern, catchment characteristics and the ground water contribution to the streamflow. The rainfall distribution, over the catchment, depends on local meteorology, large scale atmospheric circulation patterns and the geography of the catchment. It may be difficult to predict weekly streamflow accurately, by just considering the rainfall of few previous weeks, because the rainfall in current week also has substantial contribution to streamflow. This is especially true for monsoon season, when the soil is saturated, leading to insignificant infiltration. In the present study, a two-step approach is proposed for weekly streamflow prediction. In first step, current (weekly) multi-gridded rainfall is predicted with Genetic Programming (GP), based on large scale atmospheric circulation patterns, OLR and lag rainfall at grid points. Current step weekly streamflow is then predicted, by using observed gridded rainfall, up to previous weekly time step and GP predicted multi-gridded rainfall, at current weekly time step.

A single step model, for streamflow forecasting, is also developed, by using same inputs and the results are compared with the performance of two step model.

The objectives of this study are summarized as following:

(1) To develop GP-based models, for weekly multi-site (multi-gridded) rainfall prediction, based on large scale atmospheric circulation patterns with hydroclimatic teleconnection, lag multi-gridded rainfall and OLR.
(2) To develop GP-based weekly basin-scale streamflow prediction model, based on observed gridded rainfall at few previous weekly time steps, GP predicted gridded rainfall at current time step and lagged streamflow at immediate previous weekly time step.
(3) To assess the improvement in streamflow predictions due to inclusion of current step (week), GP predicted rainfall in the input set, in addition to the observed gridded rainfall up to the lag-1 weekly time step.
(4) To compare the results of the aforesaid two-step model with a single-step model, that uses lag streamflow, ENSO indices, EQUINOO indices, OLR anomaly, historical avg. rainfall and rainfall at the all grid points over last six weeks.

* Corresponding author. Tel.: +91 22 2576 7319 (Off.), +91 22 2576 8319 (Res.).
E-mail addresses: subimal@civil.iitb.ac.in, subimal.ghosh@gmail.com (S. Ghosh).

0022-1694/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhydrol.2010.10.004

The methodology adopted for streamflow prediction in the two-step model can be visualized in flowchart (Fig. 1a). The first step deals with the multi-gridded rainfall prediction, and the second step deals with the basin-scale streamflow prediction.

Similarly the methodology of a single-step model can be seen in flowchart (Fig. 1b).

The rainfall as well as streamflow-prediction models are developed by using Genetic Programming tool. The methodology is different from the methodologies mentioned in the literature (Eltahir, 1996; Piechota et al., 1997; Chiew et al., 1998; Chandimala and Zubair, 2007).

The novelty of this method lies in the inclusion of predicted current week rainfall in streamflow-prediction model that contributes to current-week streamflow, especially in monsoon season.

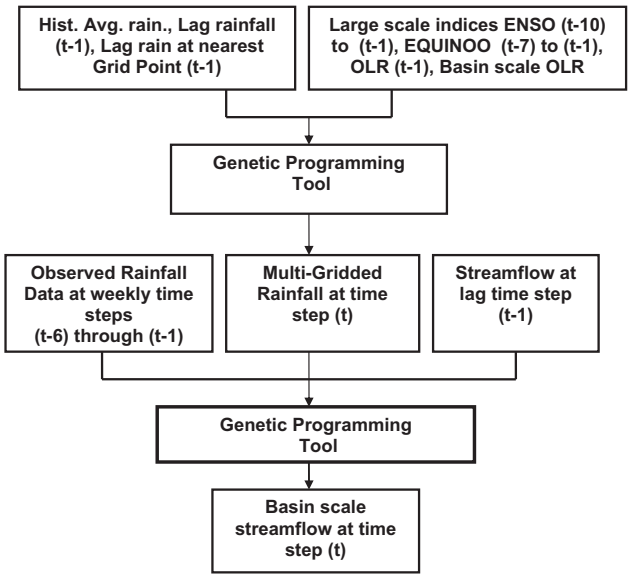

Fig. 1a. Flowchart of multi-gridded rainfall prediction followed by basin-scale streamflow prediction using Genetic Programming.

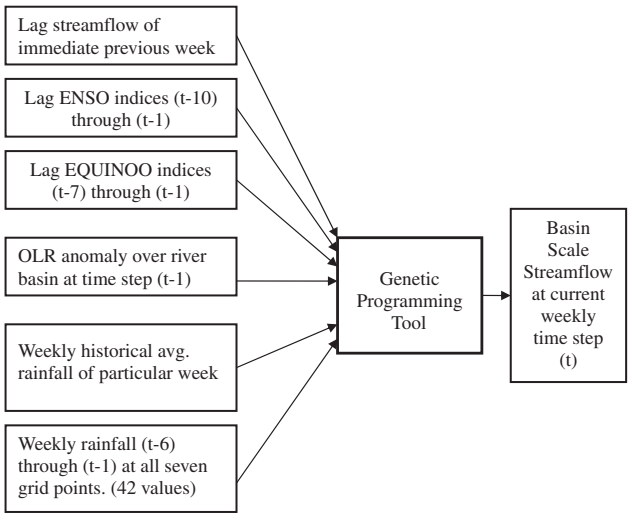

Fig. 1b. Flowchart of basin-scale streamflow prediction by single step method using Genetic Programming.

2. Influence of large scale atmospheric circulation patterns over spatio-temporal rainfall distribution

Simultaneous variations in weather and climate over widely separated regions on earth have long been noted in the meteorological literature. Such recurrent patterns are commonly referred as "teleconnections". Rainfall distribution patterns over the continents are significantly linked with the atmospheric circulation through hydroclimatic teleconnection.

It is also established that the natural variation of rainfall is linked with these large scale atmospheric circulation patterns, through hydroclimatic teleconnection (Dracup and Kahya, 1994; Eltahir, 1996; Jain and Lall, 2001; Douglas et al., 2001; Ashok et al., 2004; Marcella and Eltahir, 2008; Maity and Nagesh Kumar, 2006, 2008). Hydroclimatic teleconnection between Indian Summer Monsoon Rainfall and large scale atmospheric circulation patterns over Pacific Ocean and Indian Ocean was established in literature (Rasmusson and Carpenter, 1983; Parthasarathy et al., 1988; Krishna Kumar et al., 1999; Ashok et al., 2001; Li et al., 2001; Gadgil et al., 2003, 2004; Maity and Nagesh Kumar, 2006).

El Niño Southern Oscillation (ENSO) is the coupled ocean–atmosphere mode of variability in the tropical Pacific Ocean (Cane, 1992), whereas Indian Ocean Dipole (IOD) mode is the same over tropical Indian Ocean (Saji et al., 1999). El Niño Southern Oscillation (ENSO) is related to ocean–atmosphere interaction in the equatorial Pacific Ocean. El Niño represents anomalous warming of tropical Pacific Ocean, while La Niña represents the anomalous cooling of the oceans in the same area. This oscillation observed over the Pacific Ocean gives rise to periodic shifts in interacting winds and sea-surface exchanges. Both El Niño and La Niña are accompanied by changes in atmospheric pressures between eastern and western Pacific Ocean, known collectively as ENSO.

Another phenomenon called as Indian Ocean Dipole mode (IOD) has been observed in Indian Ocean (Saji et al., 1999), which also influences the Indian Summer Monsoon Rainfall. IOD can be described as a pattern of internal variability with anomalously low sea surface temperatures off Sumatra and high sea surface temperatures in the western Indian Ocean, with accompanying wind and precipitation anomalies (Saji et al., 1999). IOD mode has important applications on climate variability in the regions surrounding the Indian Ocean, like east Africa and Indonesia. Ashok et al. (2001) have shown that the IOD plays an important role as a modulator of the ENSO–ISMR relationship. Equatorial Indian Ocean Oscillation (EQUINOO) is the atmospheric component of the IOD mode. Equatorial zonal wind index (EQWIN) is considered as an index of EQUINOO, which is defined as negative of the anomaly of the zonal component of surface wind in the equatorial Indian Ocean Region (60°E–90°E, 2.5°S–2.5°N) (Gadgil et al., 2003, 2004). Gadgil et al. (2003, 2004) established that Indian Summer Monsoon Rainfall is not only associated with ENSO but also associated with EQUINOO.

It is observed that the strength of the hydroclimatic teleconnection decreases for smaller spatio-temporal scale. However, still, significant influence exists for sub-divisional scale for most of the geographical locations. The nature of the relationship varies across different subdivisions and different seasons (Kane, 1998; Maity and Nagesh Kumar, 2007).

Effect of El Niño Southern Oscillation (ENSO) on streamflow has been widely discussed in hydro-climatic literature. Dracup and Kahya (1994) developed a relationship between streamflow in United States of America and La Niña events. Eltahir (1996) assessed the impacts of El Niño on the flow of the Nile River in Egypt. Piechota et al. (1997) pointed out, that, relationship exists between western US streamflow and atmospheric circulation patterns, during ENSO. Chiew et al. (1998) discussed effects of ENSO on Australian rainfall, streamflow and drought. Effects of ENSO and Pacific

Inter-decadal Oscillation on water supply in the Columbia River basin have been studied by Barton and Ramirez (2004). Chandimala and Zubair (2007), attempted to predict streamflow and rainfall, based on ENSO, for water resources management, in Sri Lanka. Effects of ENSO, on streamflows, have also been studied for understanding Indian hydroclimatology (Rasmusson and Carpenter, 1983; Parthasarathy et al., 1988; Krishna Kumar et al., 1999; Ashok et al., 2001; Li et al., 2001; Gadgil et al., 2004; Maity and Nagesh Kumar, 2006). Douglas et al. (2001) attempted long-range forecasting of flows in Ganges based on ENSO information. Chowdhury and Ward (2004) studied effect of ENSO on streamflows for the Greater Ganges–Brahmaputra–Meghna Basins. Nageswara Rao (1997) studied interannual variation of monsoon rainfall in Godavari river basin, to establish its connection with ENSO. Webster and Hoyos (2004) have developed a prediction scheme for monsoon rainfall and river discharge on 15–30 days timescale in the Brahmaputra and Ganges River basins. Maity and Nagesh Kumar (2008) developed a scheme for basin-scale monthly streamflow forecasting by using the information of large scale atmospheric circulation phenomena.

When compared with the study by Maity and Nagesh Kumar (2008), which uses ENSO and EQUINOO information for monthly streamflow prediction, the present study uses OLR and multi-gridded rainfall information as an additional input to get better streamflow forecasts. The forecasts, in this present study, are also at a smaller temporal scale (weekly).

Nearly 80% of ISMR is due to the southwest monsoon in 4 months of June through September and is associated with various large-scale circulations over oceans, which regulates the amount and distribution of the rainfall, over the Indian subcontinent. However, such association is more prominently observed on large geographical scale (continental and subcontinental) when compared to small geographical scale like a river basin. Also, such association is more prominently observed for longer temporal scale (i.e. seasonal or monthly), when compared to smaller temporal scale, i.e., weekly or bi-weekly. In other words, the strength of the hydroclimatic teleconnection decreases for smaller spatio-temporal scale. However, significant influence still exists over large river basins and the nature of the relationship varies for different subdivisions and different seasons (Kane, 1998; Maity and Nagesh Kumar, 2006). The reasons behind decreased strength of hydroclimatic teleconnection for smaller spatio-temporal scale may include the local topography and weather systems. An example of such modifying factors could be the influence of cyclonic events, particularly in the vicinity of coastal areas. Local meteorological influences are also very important behind the local perturbation. Thus, apart from the large-scale circulation information, the basin-scale hydrologic variables are also supposed to be equally influenced by the local meteorological variables.

Outgoing Longwave Radiation is one of the most significant inputs, from local meteorology, for rainfall prediction. OLR is the energy leaving the earth as infrared radiation, at low energy. OLR mostly measures cloud top temperatures and gives as indication of lapse rates that prevail at low latitudes and cloud top heights. Deep clouds in these largely cumulus-convection dominated regions correspond to more intense precipitation. Thus, OLR is presented as a proxy to cumulus activity and precipitation, in the region, considered for this study.

The anomaly of OLR exhibits a negative correlation with precipitation over most of the globe (Xie and Arkin, 1998). Space and time variability analyses of the Indian Monsoon Rainfall, as inferred from satellite-derived OLR data, have been discussed by Haque and Lal (1991).

3. Data and case study

Daily gridded rainfall data at a spatial resolution of 1° latitude by 1° longitude for the period 1951–2003 is obtained from India Meteorological Department (IMD). Weekly rainfall values are calculated from these daily rainfall values for the grid points, encapsulating the upper Mahanadi River basin upstream of 'Basantpur' stream gauging station (Fig. 2). Historical average of weekly rainfall at each grid point is then computed as climatological mean rainfall at a particular grid point for a particular week. The latitudes and longitudes of grid points P1–P7, in Basin are listed in Table 1. The spatial correlations coefficient matrix for observed weekly rainfall at different grid points in Mahanadi basin is given in Table 2.

The historical rainfall data is analyzed for the grid points under consideration. Average monsoon rainfall (June–October) is found to vary between 1092 mm (at grid point P-3) and 1358 mm (at grid point P-6) over the basin under consideration. The standard deviation of monsoon rainfall amongst seven grid points is found to be 108.59 mm. The average monsoon rainfall over grid points is shown Fig. 3. The weekly variation of monsoon rainfall over Mahanadi catchment (averaged over the catchment) can be depicted in Fig. 4.

The OLR data, used in this study, for streamflow estimation, in Mahanadi basin, are collected from the region spanning 15°N–

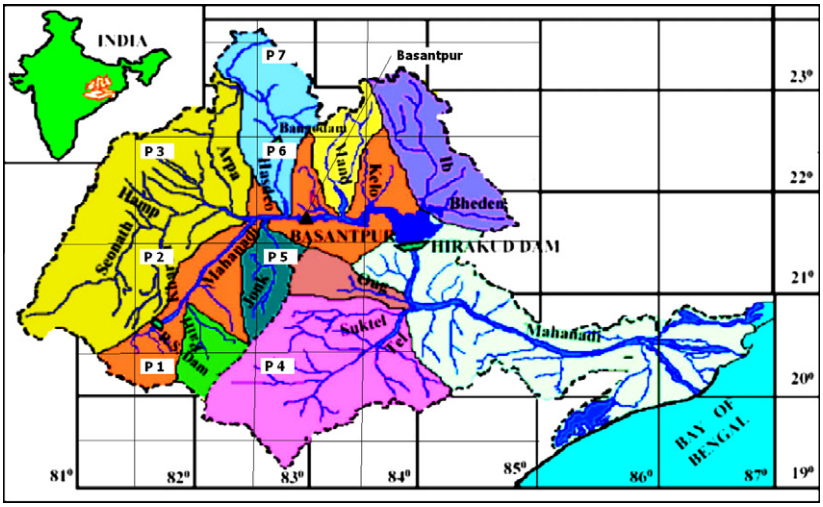

Fig. 2. One degree latitude by one degree longitude grid over Mahanadi catchment at Basantpur.

Table 1
Latitude and longitude grid points and representative geographical areas over Mahanadi catchment (refer Fig. 2).

Pt. No.	North latitude (degree)	East longitude (degree)
P1	20.5	81.5
P2	20.5	82.5
P3	21.5	81.5
P4	21.5	82.5
P5	22.5	81.5
P6	22.5	82.5
P7	23.5	82.5

Table 2
Correlations coefficient matrix for observed weekly rainfall at different grid points in Mahanadi basin.

Grid point	P-1	P-2	P-3	P-4	P-5	P-6	P-7
P-1	1.00	0.83	0.68	0.64	0.55	0.44	0.45
P-2	0.83	1.00	0.72	0.73	0.57	0.49	0.49
P-3	0.68	0.72	1.00	0.76	0.65	0.56	0.57
P-4	0.64	0.73	0.76	1.00	0.71	0.68	0.68
P-5	0.55	0.57	0.65	0.71	1.00	0.76	0.84
P-6	0.44	0.49	0.56	1.00	0.76	1.00	0.87
P-7	0.45	0.49	0.57	0.68	0.84	0.87	1.00

Correlation coefficients for observed weekly rainfall at different grid points.

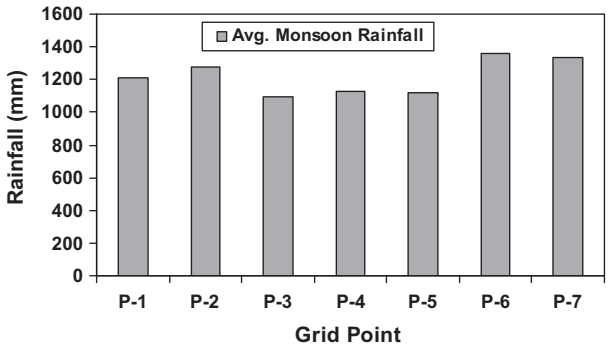

Fig. 3. Averaged monsoon rainfall over grid points (averaged for period 1951–2006).

25°N and 75°E–90°E at 2.5° latitude and longitude intervals. The daily mean values of OLR over this region, for 15 years, from 1st January 1990 to 31st December 2004, are used in this study. The extensions, beyond the catchment, are deliberately taken to capture the effect of advancing cloud systems across the basin, over a period of time. The daily mean OLR data are used to derive weekly means. The weekly mean OLR for specified region is obtained by summing the weekly grid values in a region and then dividing the sum by the number of grid points comprising the region. The OLR anomalies for the region under consideration were then computed by deducting average weekly OLR over the region from observed OLR value for the particular week. Interpolated OLR data used in this study was obtained from the NOAA web site (http://www.cdc.noaa.gov).

4. Multi-site weekly rainfall prediction methodology

It is mathematically difficult to use climate signals for the prediction of basin-scale hydrologic variables due to the inherent complexity of the climate systems. But such complex systems can be modeled by using the modern Artificial Intelligence (AI) tools, like Artificial Neural Networks (ANN), Genetic Algorithm (GA)-based evolutionary optimizer and Genetic Programming

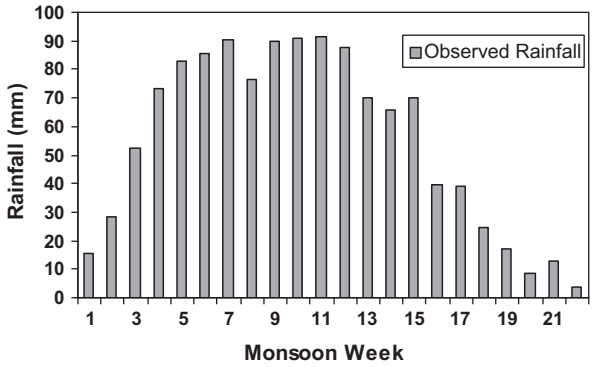

Fig. 4. Weekly monsoon rainfall over catchment (averaged over seven grid points for period 1951–2006 over basin).

(GP). Applications of ANN may be found in wide range of hydrologic studies, viz., rainfall–runoff modeling (Hsu et al., 1995; Minns and Hall, 1996), synthetic inflow generation (Raman and Sunilkumar, 1995), river flow forecasting (Dawson and Wilby, 1998; Liong et al., 2001), regional annual runoff forecasting using indices of low-frequency climatic variability (Coulibaly et al., 2000). Ozelkan and Duckstein (2001) used Fuzzy Logic (FL) based method to deal with parameter-uncertainties related to data and/or model structure. Yu et al. (2000) combined gray and fuzzy methods for rainfall forecasting. Xiong et al. (2001) used Fuzzy Logic in flood forecasting, and recommended it, as an efficient system, for flood forecasting. GA was also used in different fields of water resources engineering, e.g., rainfall–runoff modeling (Wang, 1991; Savic et al., 1999; Cheng et al., 2002), water quality models (Chau, 2002), operation of multi-reservoirs systems (Olivera and Loucks, 1997), optimal reservoir operation (Wardlaw and Sharif, 1999).

GP is a modified form of GA, applied to population of algebraic functions. GA usually operates on (coded) strings of numbers, whereas GP operates on set of algebraic functions. GP results in a regression equation, relating predictors and predictands, considering all possible combinations of algebraic operators. Koza (1992) defines GP as a domain-independent problem-solving approach, in which computer programs are evolved to solve, or approximately solved, based on the Darwinian principles of reproduction and 'survival of the fittest'.

The search procedure of a model deals with number of permutations and combinations of variables and functions in a proposed model. Any mathematical model can be represented by a tree structure. There exists a clean hierarchical structure, instead of a flat, one-dimensional string. The structure is made up of several functions that can be easily encoded using a high-level language. Any complex, nonlinear model structure can also be represented in a similar way. The genetic operators (crossover, mutation and reproduction) are performed on these trees.

It might be interesting to compare GP with the traditional approaches, such as, Autoregressive (AR) models and Neural Network (NN) approaches. ANNs do have many attractive features, but they suffer from some limitations. The difficulty in choosing the optimal network architecture and black-box nature, are the issues of concern to many researchers. In autoregressive model, only endogenous properties of the time series are used. To incorporate the external forcing, option of autoregressive model with exogenous inputs (ARX) may be selected. However, if the exogenous inputs are more, the computational complexity is a vital issue. In addition, it is linear in nature, which may not be suitable for many applications related to hydroclimatological system. On the other hand, GP has the unique feature, that it does not assume any functional form of the solution. It can optimize both the structure of the model and

its parameters. GP evolves an equation relating the output and input variables. Hence, it has the advantage of providing inherent functional relationship explicitly over techniques, such as ANN. The specialty of GP approach lies with its automatic ability to select input variables that contribute beneficially to the model and to disregard those that do not (Jayawardena et al., 2005). However, GP also has disadvantages. First, GP is a computer-intensive method and requires extensive computing power. However, owing to advent of fast computing facilities available in present days, this disadvantage can be tackled. In GP, 'possibly suitable' programs are many. This may create a dubious attitude, as it seems to be difficult to select the best (single) program. However, if it is agreed that there could be many possible line of attack to address a problem, having more than one 'possible program' would not be a skeptical issue.

This study attempts to predict weekly rainfall at individual grid points based on the lagged weekly ENSO indices, lagged weekly EQUINOO indices, rainfall at grid point at immediate previous time step, rainfall at highest correlated grid point at previous time step and previous time step OLR over the river basin.

4.1. Selection of predictors – details of correlations

It is necessary to decide exact number of weekly lags, for previous step information of ENSO indices, EQUINOO indices and rainfall, to be included in regression. Pearson's correlation coefficients are calculated between the ENSO indices and average of gridded rainfall over the basin as well as EQUINOO indices and average of gridded rainfall over the basin. The aforementioned correlations are also calculated in terms of Kendall's tau with presumption that the predictor–predictand relationship may not be linear.

The significance of these C.C. values was tested by calculating p-values of the correlations. The significance (both one-tailed and two-tailed probability values) of a Pearson correlation coefficient r, for given correlation value and the sample size are listed in the Table 3.

Pearson's correlation coefficients are also computed between the lagged and current weekly EQUINOO indices and rainfall. The significance of these C.C. values was tested by calculating p-values of the correlations. The significance (both one-tailed and two-tailed probability values) of a Pearson correlation coefficient r, for given correlation value and the sample size are listed in Table 4.

It is observed that the both one-tailed and two-tailed probability values for EN (t-10)–EN (t-1) are significant. Also, CC values, corresponding to EN (t-7)–EN (t-1), are also significant. Hence, these inputs are used for rainfall prediction as well as for streamflow prediction in this study.

It is observed from Figs. 5a and 5b that ENSO indices from (t-10) to (t-1) are significantly correlated with rainfall. Similarly it is observed from Figs. 5c and 5d that EQUINOO indices from (t-7) to (t-1) are significantly correlated with rainfall. Hence, these lagged indices are used for rainfall prediction. Similarly correlations are also computed between present rainfall and lagged rainfall (at same grid and nearby grid points). The correlations are found to be significant for lag-1 (Fig. 6). Hence, gridded rainfall at lag-1, for the same grid point and neighboring grid points, are used in rainfall prediction. The correlation values, in Figs. 5a–5d, appear to be small, in general. However, for the most complex hydroclimatic teleconnection studies, these values may be considered as good.

Table 4
p-Values indicating the significance of a Pearson correlation coefficient r between EQUINOO indices and predicted current week rainfall.

No.	EQUINOO time step (week)	Pearson's correlation coefficient with current week rainfall	Probability two tailed	Probability one tailed	Remark regarding selection of variable for prediction
1	EQ (t-8)	−0.0269	0.669	0.334	Not selected
2	EQ (t-7)	0.0804	0.203	0.102	Selected
3	EQ (t-6)	0.1223	0.052	0.026	Selected
4	EQ (t-5)	0.1206	0.056	0.028	Selected
5	EQ (t-4)	0.1553	0.014	0.007	Selected
6	EQ (t-3)	0.1499	0.017	0.009	Selected
7	EQ (t-2)	0.1456	0.021	0.010	Selected
8	EQ (t-1)	0.1275	0.043	0.022	Selected

4.2. Regression for weekly gridded rainfall estimation

The weekly rainfall at individual grid point is formulated as a function of

(i) Weekly historical avg. rainfall for present week (averaged over 1951–2003)
(ii) ENSO indices at weekly time steps: EN (t-10)–EN (t-1), i.e., 10 values
(iii) EQUINOO indices at weekly time steps: EQ (t-7)–EQ (t-1), i.e., 7 values
(iv) OLR anomaly over river basin at time step (t-1)
(v) Lag-1 Rainfall at the same grid point
(vi) Lag-1 Rainfall at nearby grid point (having highest correlation, among all grid points, with present rainfall, at station under consideration)

Thus,

$$R_t = f\{HAR_t, (EN_{t-10}, EN_{t-9} \ldots EN_{t-1}), (EQ_{t-7}, EQ_{t-6} \ldots EQ_{t-1}), OLR_{t-1}, R_{t-1}, Rn_{t-1}\} \quad (2)$$

Table 3
p-Values indicating the significance of a Pearson correlation coefficient r between ENSO indices and predicted current week rainfall.

No	ENSO time step (week)	Pearson's correlation coefficient with current week rainfall	Probability two tailed	Probability one tailed	Remark regarding selection of variable for prediction
1	EN (t-12)	-0.0128	0.8370	0.4185	Not selected
2	EN (t-11)	0.0042	0.9471	0.4735	Not selected
3	EN (t-10)	0.0131	0.8360	0.4180	Selected
4	EN (t-9)	0.0407	0.5201	0.2600	Selected
5	EN (t-8)	0.0715	0.2580	0.1290	Selected
6	EN (t-7)	0.0684	0.2790	0.1395	Selected
7	EN (t-6)	0.0726	0.2500	0.1250	Selected
8	EN (t-5)	0.0794	0.2090	0.1045	Selected
9	EN (t-4)	0.0758	0.2300	0.1150	Selected
10	EN (t-3)	0.0814	0.1970	0.0985	Selected
11	EN (t-2)	0.0691	0.2740	0.1370	Selected
12	EN (t-1)	0.0490	0.4380	0.2190	Selected

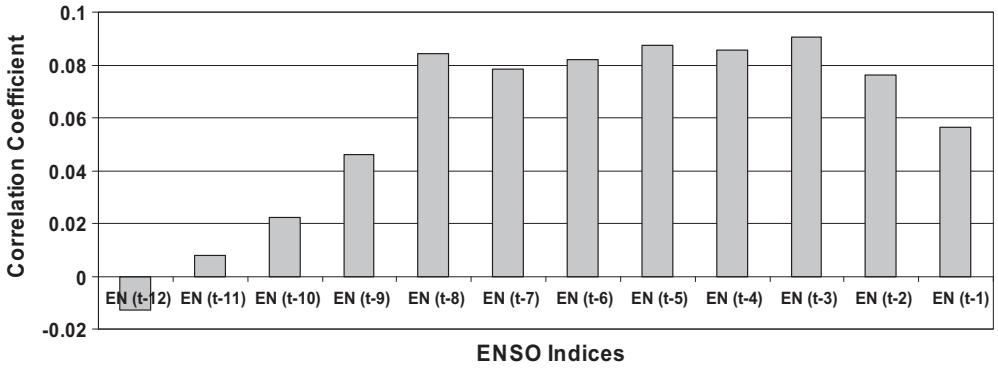

Fig. 5a. Pearson correlation coefficient between rainfall and ENSO indices with different lags.

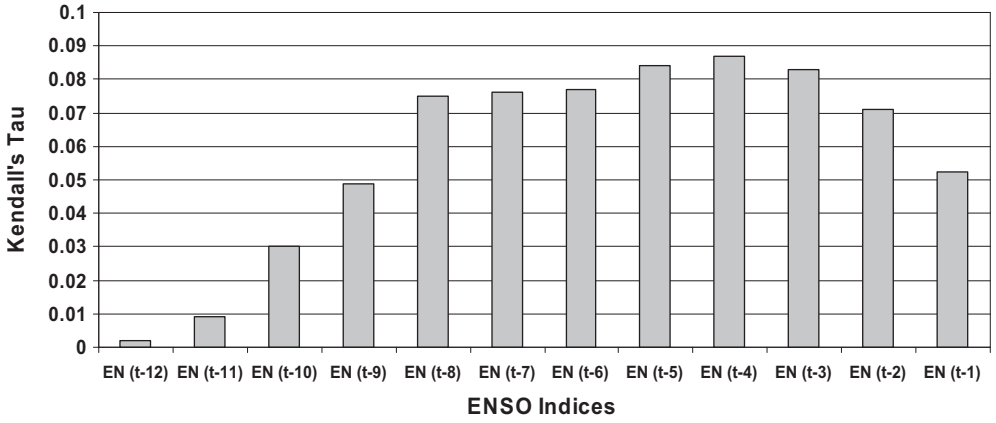

Fig. 5b. Kendall's tau between rainfall and ENSO indices with different lags.

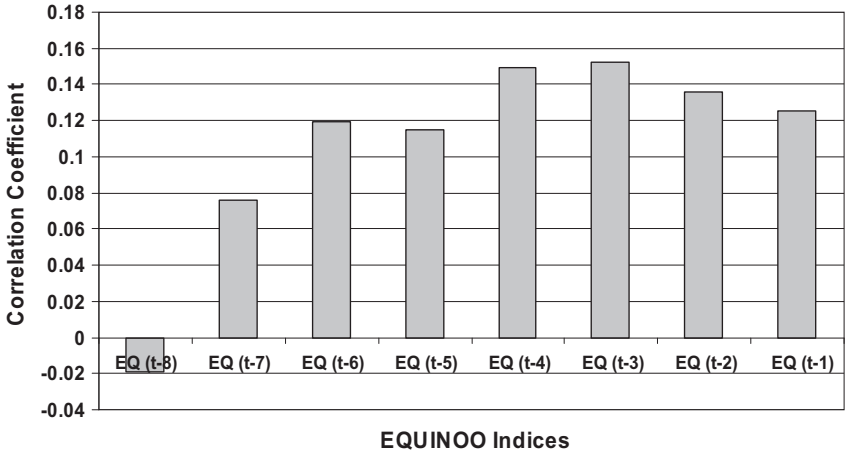

Fig. 5c. Pearson's correlation coefficient between rainfall and EQUINOO indices with different lags.

where *HAR* denotes historical weekly average Rainfall, *EN* stands for ENSO index, *EQ* stands for EQUINOO Index, *OLR* stands for Outgoing Longwave Radiation anomaly, *R* stands for weekly rainfall, *Rn* stands for rainfall at nearby grid point and *t*, *t*-1, *t*-2, etc. stand for weekly time steps.

The difference between the number of lags for ENSO and EQUINOO, considered in this formulation, can be justified as following. Earlier studies indicate that the effect of EQUINOO is more immediate than the effects of ENSO on Indian hydrologic phenomena (Maity and Nagesh Kumar, 2006). This is also convincing in the perspective of the geographical locations. As the effect of ENSO is considered up to 10 lags (approximately two and half months), the EQUINOO is considered up to lag of 7 weeks.

The data consisting of Historical average weekly rainfall, lagged rainfall, lagged ENSO indices, lagged EQUINOO indices and lagged OLR anomaly of the concerned periods are arranged in the tabular form, suitable for GP tool for the rainfall prediction. The models are trained for the period 1990–1998 and tested for 1999–2003. The analysis is limited to monsoon rainfall and monsoon streamflow only.

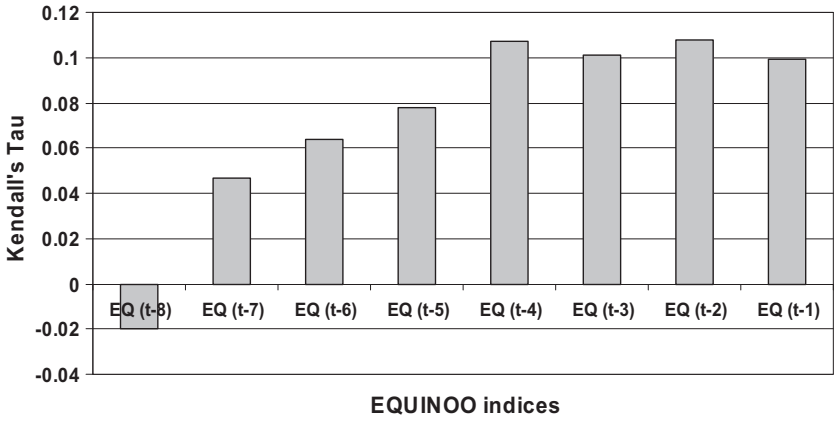

Fig. 5d. Kendall's tau between rainfall and EQUINOO indices with different lags.

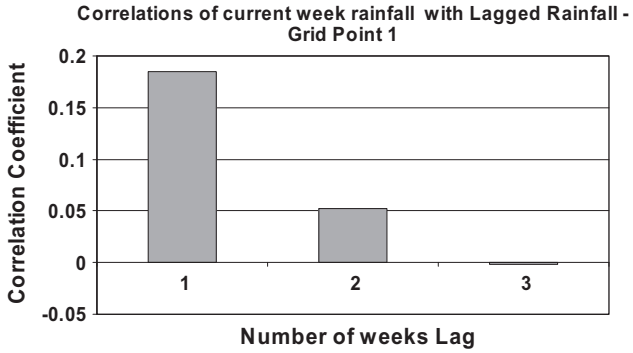

Fig. 6. Correlations of current time step point rainfall with lagged rainfall.

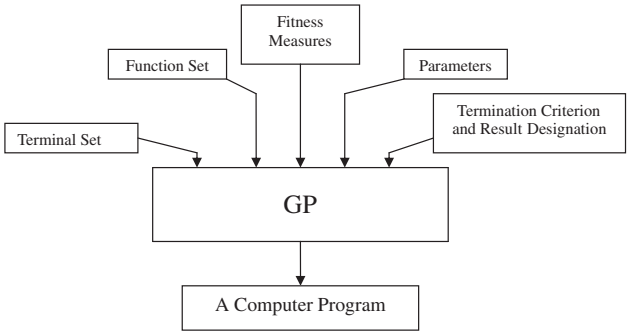

Fig. A1. Five major preparatory steps for basic version of Genetic Programming.

4.3. Genetic Programming

It is mathematically challenging to use climate signals for the prediction of basin-scale hydrologic variables, due to the inherent complexity in the climate systems. The difficulties, in modeling such complex systems, can be considerably reduced by using the modern 'Artificial Intelligence' (AI) tools. AI tools are tried nowadays for modeling complex systems. GP is basically a GA-based method, applied to a population of computer programs. While, a GA usually operates on (coded) strings of numbers, a GP operates on computer programs. The GP is similar to genetic algorithm (or rather a part of it) but unlike the later, its solution is a computer program or an equation, as against a set of numbers in GA.

GP has a unique feature, that it does not assume any functional form of the solution. It can optimize both the structure of the model and its parameters. GP evolves a computer program, representing the model, relating the output to the input variables. Hence, it has the advantage of providing inherent functional relationship, explicitly over other techniques, such as, ANN. The specialty of GP approach lies with its automatic ability to select input variables that contribute beneficially to the model and to disregard those that do not. A thorough discussion of all these concepts is covered in Koza (1992).

Application of GP needs five major preparatory steps (Koza, 1992). These five steps are (i) to select the set of terminals, (ii) to select the set of primitive functions, (iii) to decide the fitness measure, (iv) to decide parameters for controlling the run and (v) to define the method for designating the results and the criterion for terminating a run. These steps are shown in Fig. A1 – Appendix A. The choice of input variables is generally based on a priori knowledge of causal variables and physical insight into the problem being studied. If the relationship, to be modeled, is not well understood, then analytical techniques can be used. The aim of GP is to evolve a function that relates the input information to the output information, which is of the form:

$$Y = f(X^n) \tag{3}$$

where X^n is an n-dimensional input vector and Y is an output vector. In all the results reported here, Linear Genetic Programming (LGP) approach is used for formulating rainfall and streamflow-prediction models. The GP software 'Discipulus', developed by Francone (1998), Machine Learning Technologies Inc., Littleton, CO., USA, is used as a GP tool in this study. This GP tool evolves 'functions' in computer language 'C', connecting the inputs to the target output, using data sets of the input variables and output, during the 'training' process. The models are then 'tested' on the unseen data.

Genetic Programming tool is used for water resources problems by many researchers. Whigham and Crapper (2001) used Genetic Programming for rainfall–runoff modeling. Liong et al. (2001) explored GP as a flow forecasting tool. Muttil and Liong (2001) used GP for improving runoff forecasting by input variable selection. Dorado et al. (2003) used ANN and GP for prediction and modeling of the rainfall–runoff transformation for a typical urban basin. Drunpob et al. (2005) applied Genetic Programming for stream flow rate prediction in a semi-arid coastal watershed. Makkeasorn et al. (2008) compared Genetic Programming and neural network models for short-term streamflow forecasting with global climate change implications. Maity and Kashid (2010) used GP for short-term basin-scale streamflow forecasting using large-scale coupled atmospheric–oceanic circulation and local Outgoing Longwave Radiation. Details of GP algorithms are presented in Appendix A.

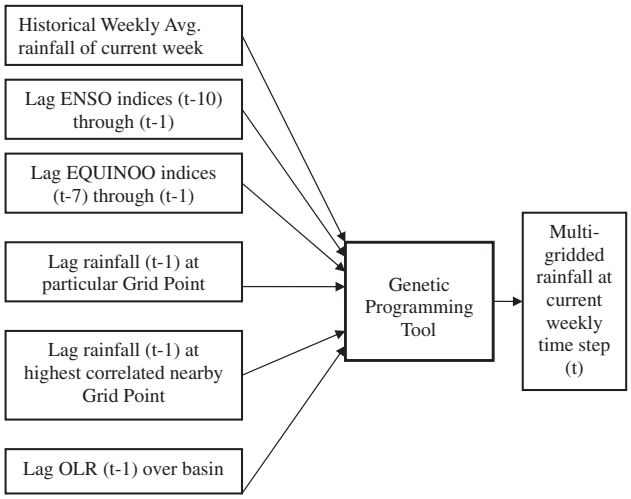

Fig. 7. Flowchart of multi-gridded rainfall prediction followed by basin-scale streamflow prediction using Genetic Programming.

4.4. Multi-site rainfall prediction

The flowchart for multi-site rainfall prediction (one grid point at a time) using Genetic Programming is shown in Fig. 7. The developed models are applied, and the results are discussed as following.

As a representative example, the plots of observed weekly rainfall and GP predicted weekly rainfall at grid point P1, during training period (1990–1998) and testing period (1999–2003) are shown in the Figs. 8a and 8b respectively. It is observed from Fig. 8a (Training) and Fig. 8b (Testing) that the weekly rainfall can be predicted with reasonable accuracy by using predictors enlisted in Section 4.2. The ENSO and EQUINOO indices, with support of OLR and lag rainfall can predict weekly rainfall with reasonable accuracy. Correct observed rainfall data at (t-6)–(t-1) are already available for streamflow prediction at the time of prediction. This approach is adopted for arriving at the better streamflow forecasts that consider the contribution of current week rainfall also (though predicted) in basin-scale streamflow prediction. Thus, the streamflow forecasts are available with one week lead time, with the prediction model, where rainfall (GP predicted) at current time step (t) is used.

The C.C. matrix is computed for grid point wise observed rainfall and GP predicted rainfall for comparison. The correlation coefficients amongst the GP predicted rainfall at different grid points are reported in Table 6. Both the C.C. matrices, i.e. for observed rainfall and GP predicted rainfall do not match perfectly. It may be considered as the limitation of the model. Such limitations for multi-site rainfall prediction are also reported by Toth et al. (2000). The limitation may be due to the fact that hydroclimatic

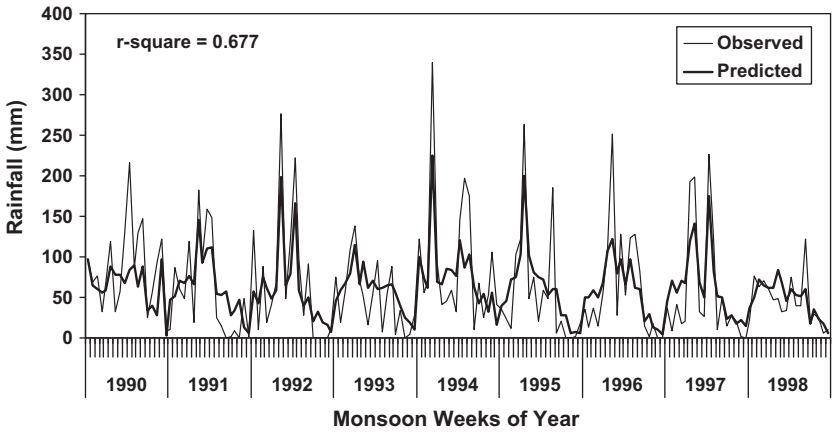

Fig. 8a. Observed and GP predicted weekly rainfall over Mahanadi basin for grid point P1 (Training).

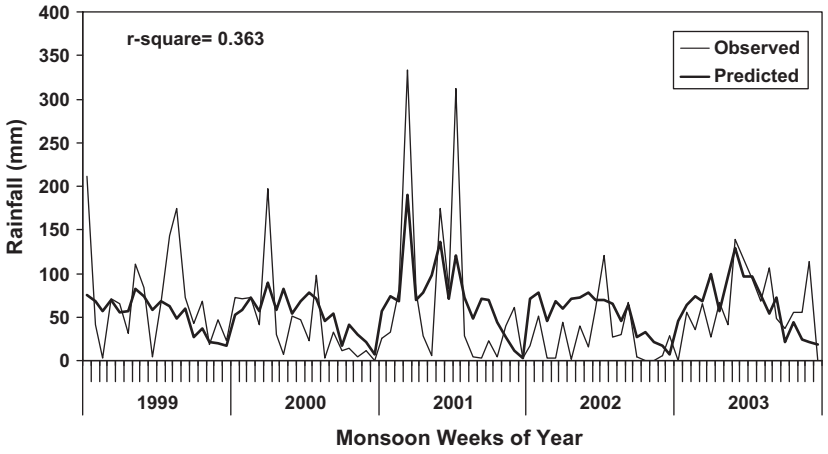

Fig. 8b. Observed and GP predicted weekly rainfall over Mahanadi basin for grid point P1 (Testing).

Table 5
The Pearson's correlation coefficients and Nash–Sutcliffe coefficient between observed and predicted rainfall for seven grid points.

Grid point	C.C. between observed and predicted rainfall (Training)	C.C. between observed and predicted rainfall (Testing)	Nash–Sutcliffe coefficient between observed and predicted rainfall (Training)	Nash–Sutcliffe coefficient between observed and predicted rainfall (Testing)
1	0.823	0.602	0.618	0.347
2	0.772	0.694	0.559	0.443
3	0.682	0.579	0.450	0.328
4	0.888	0.581	0.732	0.343
5	0.754	0.614	0.558	0.375
6	0.789	0.494	0.596	0.233
7	0.822	0.607	0.669	0.366

Note: C.C: Pearson's correlation coefficient (r).

Table 6
Correlations coefficient matrix for GP predicted weekly rainfall at different grid points in Mahanadi basin.

Grid point	P-1	P-2	P-3	P-4	P-5	P-6	P-7
P-1	1.000	0.803	0.749	0.608	0.643	0.589	0.555
P-2	0.803	1.000	0.688	0.568	0.655	0.584	0.556
P-3	0.749	0.688	1.000	0.654	0.726	0.694	0.636
P-4	0.608	0.568	0.654	1.000	0.579	0.672	0.631
P-5	0.643	0.655	0.726	0.579	1.000	0.740	0.682
P-6	0.589	0.584	0.694	0.672	0.740	1.000	0.753
P-7	0.555	0.556	0.636	0.631	0.682	0.753	1.000

Correlation coefficients for GP predicted weekly rainfall at different stations.

teleconnection is prominently experienced over large spatial scale (e.g. continental). Hence, the models are unable to reproduce spatial correlation satisfactorily.

Rainfall prediction models are developed for all seven grid points in the Mahanadi catchment. The performance of the models is assessed by calculating the Pearson's correlation coefficients between observed and computed rainfall.

The 'Nash–Sutcliffe model efficiency coefficient' is also used to assess the predictive power of rainfall prediction. It is defined as:

$$E = 1 - \frac{\sum_{t=1}^{T}\left(Q_o^t - Q_m^t\right)^2}{\sum_{t=1}^{T}\left(Q_o^t - \overline{Q_o}\right)^2} \quad (4)$$

where Q_o is observed rainfall and Q_m is modeled value of rainfall. Q_o^t is observed value of rainfall at time t and $\overline{Q}$ is mean rainfall. Nash–Sutcliffe efficiencies can range from $-\infty$ to 1. An efficiency of 1 ($E = 1$) corresponds to a perfect match of modeled discharge to the observed data. An efficiency of 0 ($E = 0$) indicates that the model predictions are as accurate as the mean of the observed data, whereas an efficiency less than zero ($E < 0$) occurs when the observed mean is a better predictor than the model. Essentially, the closer the model efficiency is to 1, the more accurate the model is. The performances of the seven rainfall prediction models in terms of Pearson's correlation coefficient and Nash–Sutcliffe coefficient are listed in Table 5.

Simultaneous rainfall prediction at all seven grid points is tried firstly using Artificial Neural Networks. But the results are not satisfactory due to too many complexities in atmospheric systems and large distances between the grid points. Also, one cannot expect same meteorological conditions over the extensive catchment area over 40,000 km^2. Hence, weekly rainfall prediction models are developed for every individual grid point separately, as discussed in Section 4.2.

The adopted Genetic Programming approach is compared with Artificial Neural Networks (ANN) and Linear Regression (LR) in terms of correlation coefficients between observed and predicted streamflow. Two layer Artificial Neural Network is chosen with 'Nguyrn-Wirdow' layer initializations function. The hyperbolic tangent sigmoid transfer function is used as the transfer function in the first layer. Linear transfer function is used in the second layer to calculate a layer's output from its net input. The network training function updates weights and bias values according to the resilient back propagation algorithm. The error performance is assessed with the Mean Squared Error (MSE) function. Table 7 presents comparative results amongst these three approaches. It is observed from Table 7 that the Genetic Programming approach gives better results when compared to Artificial Neural Networks and Linear Regression during testing phase. This underlines the efficacy of Genetic Programming in modeling such a complex system.

Table 7
Comparison of Pearson's correlation coefficient values between observed and predicted rainfall using different tools.

Grid point	Pearson's correlation coefficient					
	Genetic Programming		Artificial Neural Networks		Linear Regression	
	Training	Testing	Training	Testing	Training	Testing
1	0.823	0.602	0.745	0.255	0.537	0.515
2	0.855	0.632	0.744	0.294	0.604	0.488
3	0.682	0.579	0.743	0.208	0.611	0.561
4	0.888	0.581	0.729	0.238	0.689	0.514
5	0.754	0.614	0.786	0.304	0.645	0.611
6	0.789	0.494	0.719	0.203	0.629	0.480
7	0.822	0.607	0.671	0.272	0.607	0.551

4.5. Rainfall forecasting by excluding ENSO, EQUINOO and OLR information

The novelty of this study lies in the use of large scale atmospheric circulation pattern information (ENSO and EQUINOO) as well as Outgoing Longwave Radiation (OLR) for weekly rainfall prediction. The superiority of the model is demonstrated by comparing the models which use ENSO, EQUINOO and OLR information with those which do not use the ENSO, EQUINOO and OLR information. Table 8 compares the r^2 values during training and testing, for

Table 8
Comparison of GP models which do not use the ENSO, EQUINOO and OLR information (uses average streamflow for present week) with those which use ENSO, EQUINOO and OLR information.

Grid point	Analysis without using ENSO EQUINOO and OLR information		Analysis by using ENSO, EQUINOO and OLR information	
	r^2 (Training)	r^2 (Testing)	r^2 (Training)	r^2 (Testing)
1	0.522	0.333	0.677	0.362
2	0.487	0.348	0.731	0.399
3	0.438	0.286	0.465	0.335
4	0.653	0.178	0.788	0.337
5	0.469	0.322	0.569	0.377
6	0.507	0.216	0.623	0.244
7	0.62	0.290	0.676	0.369

the GP models, which do not use ENSO and EQUINOO information, with those, using ENSO and EQUINOO information.

It is observed from results (Table 8) that the *r*-square values between observed and predicted rainfall during training and testing show improvements, when the ENSO, EQUINOO and OLR information are used. This indicates the usefulness of global as well as local meteorological inputs for rainfall prediction.

5. Streamflow prediction by rainfall–runoff modeling approach

Genetic Programming is used to translate the gridded rainfall information into streamflow. Auto-correlations are normally found to be significant in streamflow studies. Hence, streamflow in immediate previous time step is also considered as input in the present study. The basin-scale streamflow prediction is performed using data from 1990 to 2003, among which, monsoon rainfall data of years 1990–1998 are used for training purpose. The data from 1999 to 2003 are used for testing purpose.

Three different analyses are carried out by using three alternative methodologies which are described as follows.

(i) First, the rainfall–runoff models are developed for weekly streamflow (SF_t) prediction, with rainfall information as, R_O (t-6)–R_O (t-1), and streamflow during immediate previous week.

$$SF_t = f\{(SF_{t-1}), (R_O1)_{t-6}, (R_O2)_{t-6}, (R_O3)_{t-6}, \ldots, (R_O7)_{t-6}), ((R_O1)_{t-5}, (R_O2)_{t-5}, (R_O3)_{t-5}, \ldots, (R_O7)_{t-5}), \ldots ((R_O1)_{t-1}, (R_O2)_{t-1}, (R_O3)_{t-1}, \ldots, (R_O7)_{t-1})\}, \quad (5)$$

where SF is streamflow, $R_O1, R_O2, \ldots, R_O7$ are observed rainfall values at Grid Points P-1, P-2, …, P7, *t, t-1, t-2 , …,t*-6 are weekly time steps.

(ii) Rainfall–runoff models are developed with weekly area weighted observed rainfall for (t-6) to (t) time steps, for streamflow (SF_t) prediction.

$$SF_t = f\{(SF_{t-1})((R_O1)_{t-6}, (R_O2)_{t-6}, (R_O3)_{t-6}, \ldots, R_O7)_{t-6}), ((R_O1)_{t-5}, (R_O2)_{t-5}, (R_O3)_{t-5}, \ldots, R_O7)_{t-5}), \ldots ((R_O1)_t, (R_O2)_t, (R_O3)_t, \ldots, R_O7)_t)\}, \quad (6)$$

(iii) Rainfall–runoff models with observed rainfall for (t-6)–(t-1) time steps and GP predicted rainfall at time step (t) [$R_p(t)$] for weekly streamflow prediction. Thus, the equation can be written as:

$$SF_t = f\{(SF_{t-1}), (R_O1)_{t-6}, (R_O2)_{t-6}, (R_O3)_{t-6}, \ldots, R_O7)_{t-6}), ((R_O1)_{t-5}, (R_O2)_{t-5}, (R_O3)_{t-5}, \ldots, R_O7)_{t-5}), \ldots ((R_P1)_t, (R_P2)_t, (R_P3)_t, \ldots, R_P7)_t)\}, \quad (7)$$

The developed models are used for streamflow prediction with observed rainfall up to (t-1) time step and GP predicted rainfall at time step (t). The advantages of this approach are first, it considers current step rainfall also in streamflow prediction, for better accuracy; second, the forecasts are available in one week lead time.

6. Streamflow prediction by single-step method

The problem of weekly streamflow forecasting is solved with single-step method, for comparison, by directly relating large-scale circulation and local meteorological information with streamflow.

The revised model uses

(i) Lagged streamflow in immediate previous week
(ii) ENSO indices at weekly time steps: (t-10)–(t-1), i.e., 10 values
(iii) EQUINOO indices at weekly time steps: (t-7)–(t-1), i.e., 7 values
(iv) OLR anomaly over river basin during time step (t-1)
(v) Weekly historical avg. rainfall during particular week (averaged over 1951–2003)
(vi) Rainfall at the same grid point in over last 6 weeks.

Thus, the weekly rainfall at individual grid point is formulated as a function of

$$SF_t = f\{\{SF_{t-1}, (EN_{t-10}, EN_{t-9}, \ldots, EN_{t-1}), (EQ_{t-7}, EQ_{t-6} \ldots, EQ_{t-1}), OLR_{t-1}\}, (HAR_t1, HAR_t2, HAR_t3, HAR_t4, HAR_t5, HAR_t6, HAR_t7), ((R_O1)_{t-6}, (R_O2)_{t-6}, (R_O3)_{t-6}, \ldots, R_O7)_{t-6}), ((R_O1)_{t-5}, (R_O2)_{t-5}, (R_O3)_{t-5}, \ldots, R_O7)_{t-5}), \ldots ((R_O1)_{t-1}, (R_O2)_{t-1}, (R_O3)_{t-1}, \ldots, (R_O7)_{t-1})\} \quad (8)$$

where SF is streamflow, *HAR* stands for historical weekly average Rainfall at particular grid point (1, 2, 3, etc.), *EN* stands for ENSO index, *EQ* stands for EQUINOO Index, *OLR* stands for Outgoing Longwave Radiation anomaly, R_O stands for weekly observed rainfall, and t, t-1, t-2, etc. stand for weekly time steps. $R_O1, R_O2, \ldots, R_O7$ are observed rainfall values at Grid Points P1, P2, …, P7. t, t-1, t-2, …, t-6 are weekly time steps.

7. Results and Discussion

First, the performance of rainfall–runoff models, developed by using weekly observed rainfall during weeks, (t-6)–(t-1) are evaluated in terms of correlation coefficient. The correlation coefficients between observed streamflow and GP-predicted streamflow are calculated. They are 0.715 during training and 0.741 during testing. (r^2 = 0.514 during training and r^2 = 0.555 during testing). The plots of Observed and GP-Predicted Streamflow, using this approach, during training and testing are presented in Figs. 9a and 9b respectively.

The rainfall–runoff models, presented in Eq. (6), are then used for prediction of streamflow, with weekly area weighted observed rainfall for time steps (t-6)–(t) as inputs. This combination is observed to predict streamflows satisfactorily, which can be depicted from correlation coefficient of 0.931 during training and 0.817 during testing (r^2 = 0.867 during training and r^2 = 0.669 during testing). The improvement in results for approach 2 (Eq. (6)), over approach 1 (Eq. (6)), is due to the inclusion of observed rainfall at current time step (t) for streamflow prediction. This also underlines the utility of rainfall value at time step (t) for streamflow prediction. The plots of Observed and GP Predicted Streamflow, by this approach, during training and testing, are presented in Figs. 10a and 10b, respectively.

However, rainfall at present time step cannot be used in one week lead time. Hence, it is decided to use GP-predicted gridded rainfall values, at time step (t) in place of observed rainfall values, at present time step. In the third model (Eq. (7)), rainfall–runoff relationship is developed, with observed rainfall from (t-6) to (t-1) and GP predicted rainfall, during time period (t). The correlation coefficient between observed and predicted streamflow are found to be 0.812 (r^2 = 0.667). The slight reduction, in this value when compared to approach 2, may be due to the deviations of GP-predicted rainfall, with respect to observed rainfall, at time step (t). The plots of observed and GP-predicted streamflow by this third approach (Eq. (7)) are presented in Fig. 10c. The predictions are also evaluated based on the Nash–Sutcliffe model efficiency coefficients. The performances of above three models in terms of Pearson's correlation coefficients and Nash–Sutcliffe model efficiency coefficients are tabulated in Table 9.

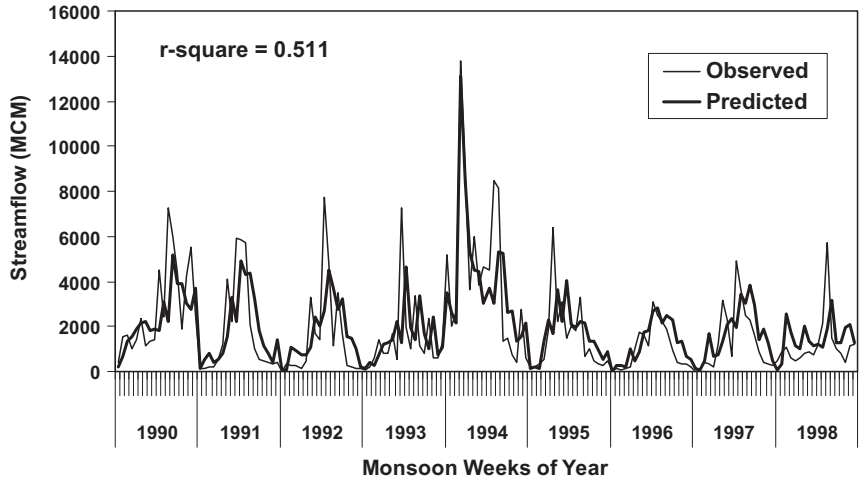

Fig. 9a. Observed and GP predicted streamflow by rainfall–runoff approach with input as observed rainfall at (t-6)–(t-1) time steps (Training).

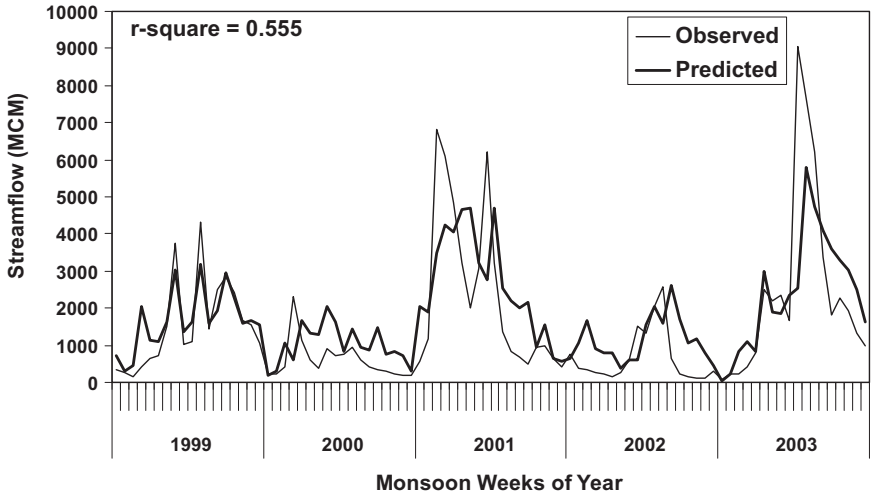

Fig. 9b. Observed and GP predicted streamflow by rainfall–runoff approach with input as observed rainfall at (t-6)–(t-1) time steps (Testing).

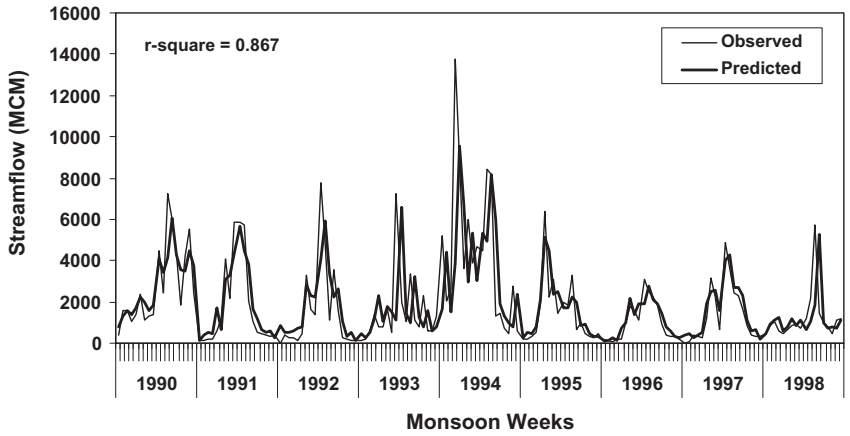

Fig. 10a. Observed and GP predicted streamflow by rainfall–runoff approach with input as observed rainfall at time steps (t-6)–(t) (Training).

Thus, it can be concluded that the two-step methodology of weekly streamflow prediction, with historical average rainfall of current time step, observed rainfall up to (t-1) time step and GP predicted rainfall at current time step (based on ENSO, EQUINOO and Lag-1 rainfall at grid point and Lag-1 OLR), gives reasonably accurate basin-scale streamflow forecasts, with 1 week lead time. The forecasts can be certainly useful for basin-scale real time water management. Similar models can be developed for basin-scale

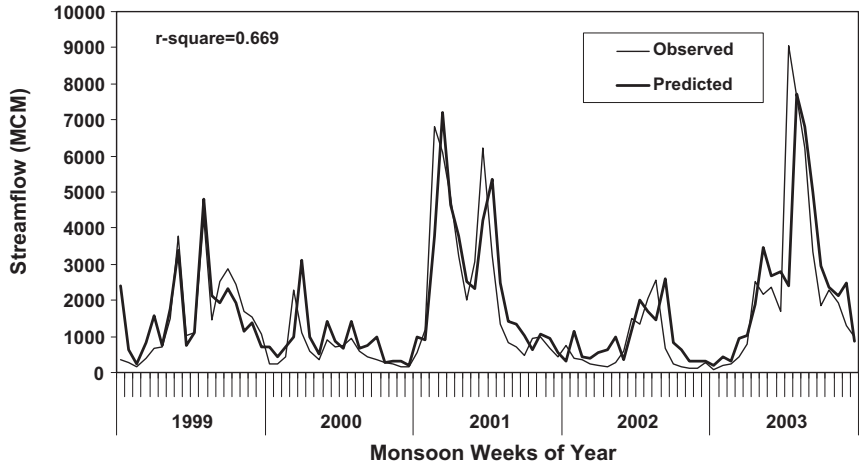

Fig. 10b. Observed and GP predicted streamflow by rainfall–runoff approach with input as observed rainfall at time steps (t-6)–(t) (Testing).

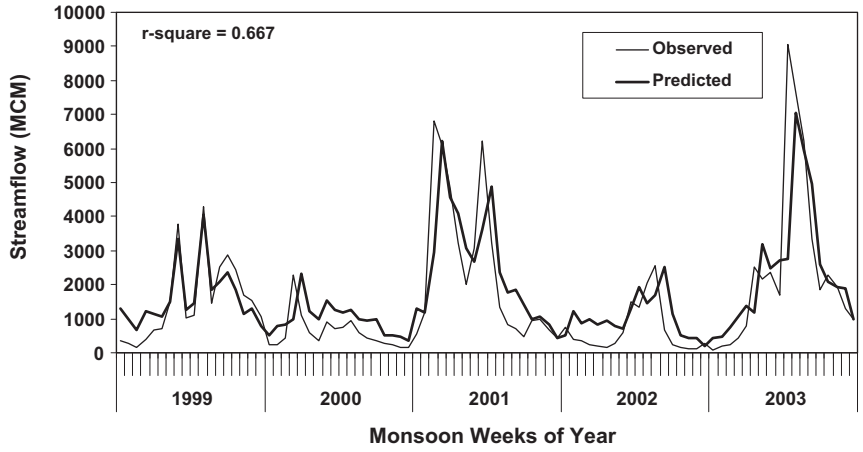

Fig. 10c. Observed and GP predicted streamflow by rainfall–runoff approach with input as observed rainfall at (t-6)–(t-1) time steps and GP predicted rainfall at time step (t) (Testing).

Table 9
Comparison of streamflow prediction approaches.

Approach no.	Streamflow is a function of the following	r^2 Training	r^2 Testing	N.S. Coeff. Training	N.S. Coeff. Testing
1	Observed gridded rainfall (t-6)–(t-1) time steps, Streamflow (t-1)	0.715	0.741	0.509	0.533
2	SF (t-1), observed rainfall (t-6)–(t)	0.867	0.669	0.775	0.660
3	SF (t-1), observed rainfall (t-6)–(t-1) GP predicted rainfall (t)		0.667	0.775 Same as (2) as using same model	0.648

Note: C.C.: Pearson's correlation coefficient. N.S. Coeff.: Nash–Sutcliffe model efficiency coefficient.

Table 10
Results of a single step streamflow prediction model.

Approach	Streamflow is a function of the following	r^2 Training	r^2 Testing	N.S. Coeff. Training	N.S. Coeff. Testing
1	Streamflow (t-1), ENSO (t-12)–(t-1), EQUINOO (t-12)–(t-1), OLR (t-2)–(t-1), historical average rainfall at each grid point, lagged gridded rainfall at each grid point, lagged gridded rainfall at highest correlated grid point.	0.688	0.656	0.677	0.642

Table 11
Comparison of Pearson's correlation coefficient values between observed and predicted streamflow using different tools.

Grid point	Pearson's correlation coefficient					
	Genetic Programming		Artificial Neural Networks		Linear Regression	
	Training	Testing	Training	Testing	Training	Testing
1	0.829	0.810	0.876	0.21	0.727	0.702

streamflow prediction for other river basins, where hydroclimatic teleconnection is prominently noticed.

Results of the single-step methodology can be discussed as following. The r^2 values between observed and predicted streamflow during training and testing are found to be 0.688 and 0.656, respectively, when compared to 0.867 during training and 0.669 during testing, for a two-step model (Refer Table 10 for GP results and Table 11 for the comparison).

It is observed that the results of single step model are inferior, for both training and testing. The results can also be compared in terms of simulating peak streamflows. (Figs. 10c,11a and 11b). Though the single-step model shows comparable r^2 value during testing of GP models, the plots of observed and predicted streamflow for two-step model show that the peak streamflows are better simulated by the two-step model. This is due to the most natural rainfall–runoff approach, adopted for streamflow prediction, in the two-step methodology.

It is well understood from the analysis that the local meteorological information over the catchment is as important as large-scale circulation pattern information for rainfall prediction. Local meteorological information, in form of Outgoing Longwave Radiation (OLR), is thus, included in rainfall prediction models at all seven grid points.

The reasons behind using weekly ENSO index, in rainfall as well as streamflow forecasting, is discussed here. The strength of easterly trade winds and the amount of moisture transfer are largely influenced by the sea surface temperature anomalies and associated pressure anomalies over tropical Pacific Ocean. The El Niño Southern Oscillation Index (ENSO index) happens to be the indicator of these activities observed over tropical Pacific Ocean. 'El Niño Southern Oscillation' can show three conditions viz. 'El Niño' conditions, 'Normal' conditions and 'La Niña' conditions. The streamflow forecasting models perform well in all the times, i.e. during 'El Niño' conditions, 'Normal' conditions and 'La Niña' conditions. Furthermore, 'El Niño' and 'La Niña' can also be 'weak', 'moderate' or 'strong', depending upon the 'Oceanic Niño Index' (ONI) used by NOAA.

El Niño conditions are said to be developed when SST anomaly remains on the negative side of −0.5 over seven consecutive months, whereas La Niña conditions are said to be developed when SST anomaly on positive side of +0.5 over seven consecutive months. The conditions in between are called as 'Normal conditions'. For Indian Summer Monsoon, the sea surface temperature anomalies from March to September are most important, as monsoon rains extends from June to middle of October in India. It can be interesting to know the ENSO status over the period of analysis. Accordingly the yearly ENSO conditions are listed as following: 1990 – normal conditions, 1991 – strong El Niño, 1992 – normal conditions, 1993 – normal conditions, 1994 – moderate El Niño, 1995 – weak La Niña, 1996 – normal conditions, 1997 – strong El Niño, 1998 – moderate La Niña, 1999 – moderate La Niña, 2000 – weak La Niña, 2001 – normal conditions, 2002 – moderate El Niño, 2003 – normal conditions (Source: http://www.cpc.noaa.gov/products/analysis_monitoring/ensostuff/ensoyears.shtml).

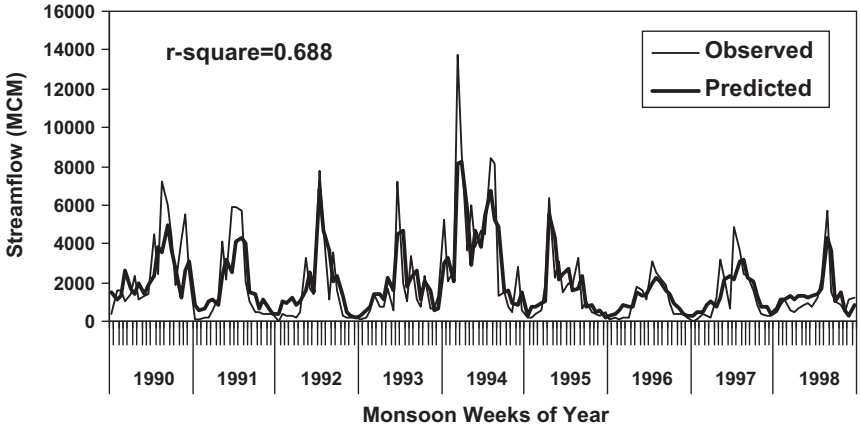

Fig. 11a. Observed and GP predicted streamflow by single step approach (Training).

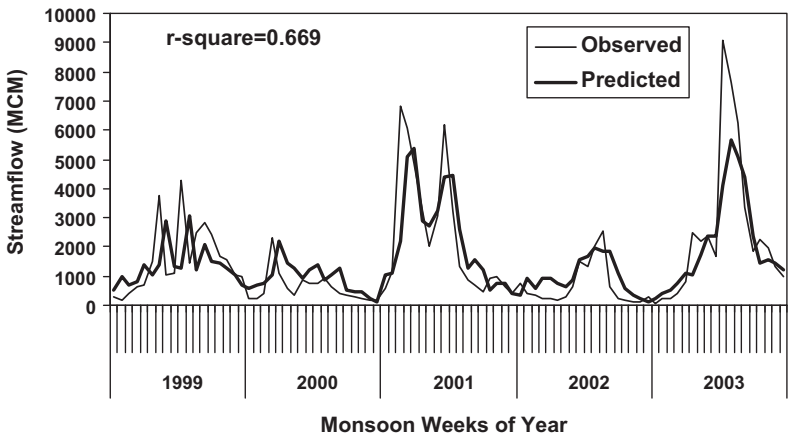

Fig. 11b. Observed and GP predicted streamflow by single step approach (Testing).

The models in this study use weekly ENSO indices of 12 immediate previous weeks (12 values) and EQUINOO indices of 7 immediate previous weeks (7 values) as inputs for weekly streamflow prediction. It should be noted here that several weeks of lags are used to incorporate evolutionary trends in the data of the input variables.

It is observed that lagged negative values of ENSO indices lead to below normal rainfall and streamflow. On the other hand, the lagged positive ENSO indices lead to above normal rainfall and streamflow. The performances of model in monsoon weeks of years 1999, 2000, 2001, 2002 and 2003 are observed in this study. Out of 90 weekly streamflow data, in testing, it is observed that 65 are correct. This shows the efficacy of ENSO index for rainfall and streamflow prediction.

The studies by Eltahir (1996), Piechota et al. (1997) and Chiew et al. (1998) use ENSO as the principal indicator of streamflow variation, whereas the present study includes EQUINOO from Equatorial Indian Ocean and basin-scale OLR for rainfall and streamflow prediction. Inclusion of EQUINOO and local meteorological information (OLR) makes this study different from the earlier studies.

The rainfall prediction models as well as streamflow-prediction models, developed in this study, use large number of predictors. The selection of predictors, as well as the improvement in predictive skill, from each variable, in such cases can benefit from the use of nonlinear dependence measures, which are robust to noise and short length of the data. However, this can be the future scope of the present study.

8. Concluding remarks

The information of large scale atmospheric circulation patterns viz. El Niño Southern Oscillation (ENSO) and Equatorial Indian Ocean Oscillation (EQUINOO), with support of lagged rainfall information at every individual grid point and basin-scale OLR anomaly are successfully used for prediction of weekly gridded monsoon rainfall, in Mahanadi catchment with reasonable accuracy.

Results of this study show that the inclusion of GP-predicted rainfall, at current time step (t), as input, for weekly streamflow prediction model, gives better basin-scale streamflow forecasts, when compared to streamflow prediction based on observed gridded rainfall up to the last weekly, i.e. (t-1) time step.

GP-derived rainfall–runoff models that use observed gridded rainfall up to weekly time step (t-1) and GP predicted gridded rainfall at weekly time step (t) give better streamflow forecasts than those models using rainfall up to (t-1) time step.

The blending large-scale circulation information in form of ENSO and EQUINOO indices, local meteorological information in form of OLR and lagged rainfall at grid points can be advantageous for the basin-scale streamflow forecasting.

The efficacy of Genetic Programming approach can be realized through experience of modeling of the most complex hydrometeorological systems analyzed in this study.

Appendix A. Genetic Programming approach

This appendix describes the Genetic Programming (GP) approach, applied in the studies, reported in this paper. GP is basically a genetic algorithm (GA) applied to a population of computer programs. While a GA usually operates on (coded) strings of numbers, a GP operates on computer programs. The GP is similar to genetic algorithm (or rather a part of it) but unlike the latter, its solution is a computer program or an equation, as against a set of numbers in GA. Koza (1992) defines GP as a domain independent problem-solving approach in which computer programs are evolved to solve, or approximately solve, problems based on the Darwinian principles of reproduction and 'survival of the fittest'.

Genetic Programming starts with solving a problem, by creating massive amount of simple random functions, in a population pool. These simple parent functions mate and reproduce massive amount of children offspring functions. Each offspring function is measured against the training data. Those offspring functions that closely match the training data may be kept and be allowed to reproduce, while some of the poor-fitted offspring functions would be terminated. The selected offspring functions, determined by their fitness, can reproduce another generation of grandchildren functions. Each grandchildren function may be tested against the training input data for its fitness. The good-fitted grandchildren functions may be kept and used to reproduce the next generation. Some low-fitting grandchildren functions may be terminated. This population of functions is progressively evolved over a series of generations. The search for the best result in the evolutionary process involves applying the principle of survival of the fittest. The GP can reproduce and terminate millions of function over thousands of generations to find the strongest function that fits the training input data the most. Regression models generated from the GP are free from any particular model structure (Chang and Chen, 2000).

The glass box characteristic of GP reveals structures of the regression models, which is the significant advantage of the GP over black box approaches such as neural networks. The GP model could be the best solver for searching highly nonlinear spaces for global optima via adaptive strategies.

The three important operators used in GP are crossover, reproduction, and mutation. These can be described in brief which are as follows. According to Koza (1992), the operator 'crossover' is mainly responsible for the genetic diversity in the population of programs. Similar to GA, crossover (a binary operator) operates on two programs in GP and produces two child programs. These new programs become part of the next generation of programs to be evaluated. The operation 'Reproduction' is performed by simply copying a selected member from the current generation to the next generation. Mutation becomes an important operator in genetic algorithms, which provides diversity to the population. However, as per Koza (1992), mutation is relatively unimportant in the Genetic Programming, because the dynamic sizes and shapes of the individuals in the population already provide diversity. Mutation can be rather considered as a variation on the crossover operation in GP. The flowchart of Genetic Programming methodology is shown in Fig. A2.

Application of GP needs five major preparatory steps (Koza, 1992). These five steps are (i) to select the set of terminals, (ii) to select the set of primitive functions, (iii) to decide the fitness measure, (iv) to decide parameters for controlling the run, and (v) to define the method for designating the results and the criterion for terminating a run. A flow chart showing five major preparatory steps involved in basic version of GP is shown in Fig. A1. The choice of input variables is generally based on a priori knowledge of casual variables and physical insight into the problem being studied. If the relationship to be modeled is not well understood, then analytical techniques can be used. The aim of GP is to evolve a function that relates the input information to the output information, which is of the form:

$$Y^m = f(X^n) \tag{9}$$

Where X^n, an n-dimensional input, is vector, and Y^m is an m-dimensional output vector. For example, for the weekly streamflow prediction problem, the input vector may consist of lag streamflow

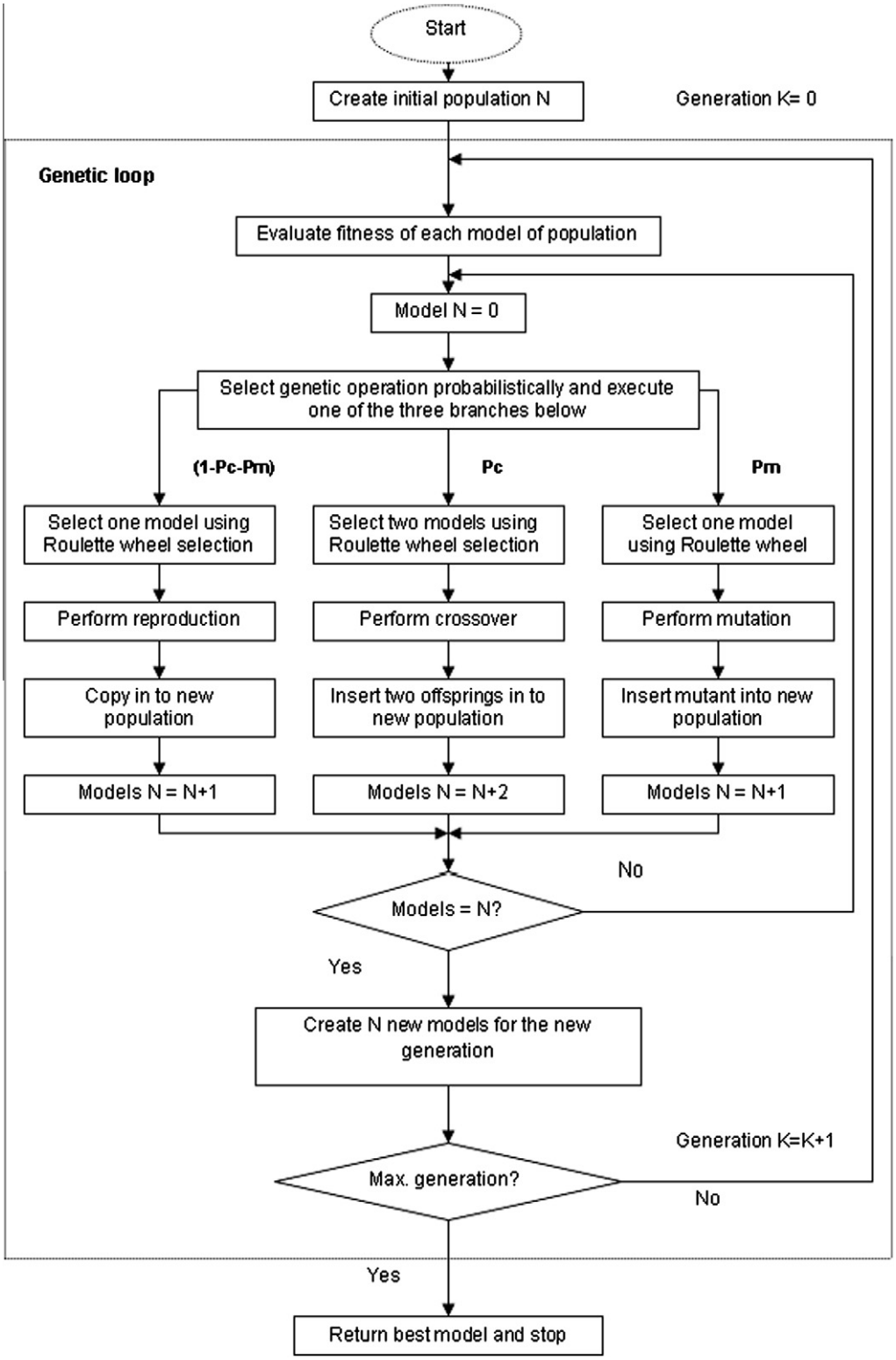

Fig. A2. Flowchart for Genetic Programming (Hong and Bhamidimarri, 2003).

(of previous week), lagged ENSO and EQUINOO indices of certain number weeks and OLR anomaly of previous week. The output can be of streamflow at current weekly time step.

The implementation of GP in this work is done through software Discipulus (Francone, 1998) that is based on an extension of the originally envisaged GP called Linear Genetic Programming (LGP). It evolves sequences of instructions from an imperative programming language or machine language. The LGP expresses instructions in a line-by-line mode. The term "linear" in Linear Genetic Programming refers to the structure of the (imperative) program representation. It does not stand for functional genetic programs that are restricted to a linear list of nodes only. Genetic

programs normally represent highly nonlinear solutions in this meaning (Brameier, 2004).

References

Ashok, K., Guan, Z., Yamagata, T., 2001. Impact of Indian Ocean dipole on the relationship between the Indian monsoon rainfall and ENSO. Geophysics Research Letters 28, 4499–4502.

Ashok, K., Guan, Z., Saji, N., Yamagata, T., 2004. Individual and combined effect of ENSO and Indian Ocean dipole on the Indian summer monsoon. Journal of Climate 17 (16), 3141–3155. doi:10.1175/1520-0442(2004)017<3141:IACIOE>2.0.CO;2.

Barton, S.B., Ramirez, J.A., 2004. Effects of El Niño Southern Oscillation and Pacific Interdecadal Oscillation on water supply in the Columbia River basin. Journal of Water Resources Planning and Management 130 (4), 281–289.

Brameier, M., 2004. On Linear Genetic Programming. PhD Thesis, Fachbereich Informatik, Universit"at Dortmund, Germany.

Cane, M.A., 1992. Tropical Pacific ENSO Models: ENSO as a mode of couples system. In: Trenberth, K.E. (Ed.), Climate System Modeling. Cambridge University Press, UK.

Chandimala, J., Zubair, L., 2007. Predictability of stream flow and rainfall based on ENSO for water resources management in Sri Lanka. Journal of Hydrology 335, 303–312.

Chang, N.B., Chen, W.C., 2000. Prediction of PCDDs/PCDFs emissions from municipal incinerators by genetic programming and neural network modeling. Waste Management & Research 18, 341–351.

Chau, K.W., 2002. Calibration of flow and water quality modeling using genetic algorithms. Lecture Notes in Artificial Intelligence 2557, 720.

Cheng, C.T., Ou, C.P., Chau, K.W., 2002. Combining a fuzzy optimal model with a genetic algorithm to solve multi-objective rainfall–runoff model calibration. Journal of Hydrology 268 (3), 72–86.

Chiew, F.H.S., Piechota, T.C., Dracup, J.A., McMahon, T.A., 1998. El Niño/Southern Oscillation and Australian rainfall, streamflow and drought: links and potential for forecasting. Journal of Hydrology 204, 138–149.

Chowdhury, M.R., Ward, N., 2004. Hydro-metrological variability in the greater Ganges–Brahmaputra–Meghna basins. International Journal of Climatology 24, 1495–1508.

Coulibaly, P., Anctil, F.F., Rasmussen, P., Bobee, B., 2000. A recurrent neural networks approach using indices of low-frequency climatic variability to forecast regional annual runoff. Hydrological Processes 14, 2755–2777.

Dawson, C.W., Wilby, R., 1998. An artificial neural network approach to rainfall–runoff modeling. Hydrological Sciences Journal 43 (1), 47–66.

Dorado, J., Rabunal, J.R., Pazos, A., Rivero, D., Santos, A., Puertas, J., 2003. Prediction and modelling of the rainfall–runoff transformation of a typical urban basin using ANN and GP. Applied Artificial Intelligence 17, 329–343.

Douglas, W.W., Wasimi, S.A., Islam, S., 2001. The El Niño Southern Oscillation and long-range forecasting of flows in Ganges. International Journal of Climatology 21, 77–87.

Dracup, J.A., Kahya, E., 1994. The relat0069onship between US streamflow and La Niña events. Water Resources Research 30 (7), 2133–2141.

Drunpob, A., Chang, N.B., Beaman M., 2005. Stream flow rate prediction using genetic programming model in a semi-arid coastal watershed. In: Proceedings of EWRI 2005, ASCE.

Eltahir, E.A.B., 1996. El Niño and the natural variability in the flow of the Nile River. Water Resources Research 32 (I), 131–137.

Francone, F.D., 1998. Discipulus Owner's Manual. Machine Learning Technologies, Inc., Littleton, Colorado.

Gadgil, S., Vinayachandran, P.N., Francis, P.A., 2003. Droughts of the Indian summer monsoon: role of clouds over the Indian Ocean. Current Science 85 (2), 1713–1719.

Gadgil, S., Vinayachandran, P.N., Francis, P.A., Gadgil, S., 2004. Extremes of the Indian summer monsoon rainfall. ENSO and equatorial Indian Ocean Oscillation. Geographical Research Letter 31, L12213. doi:10.1029/2004GLO19733.

Haque, M.A., Lal, M., 1991. Space and time variability analyses of the Indian monsoon rainfall as inferred from satellite-derived OLR data. Climate Research 1, 187–197.

Hong, Y.S., Bhamidimarri, R., 2003. Evolutionary self-organising modelling of a municipal wastewater treatment plant. Water Research 37, 1199–1212. doi:10.1016/S0043-1354(02)00493-1.

Hsu, K.L., Gupta, H.V., Sorooshian, S., 1995. Artificial neural network modeling of the rainfall–runoff process. Water Resources Research 31 (10), 2517–2530.

Jain, S., Lall, U., 2001. Floods in a changing climate: does the past represent the future? Water Resources Research 37 (12), 3193–3205.

Jayawardena, A.W., Muttil, N., Fernando, T.M.K.G., 2005. Rainfall–Runoff Modeling Using Genetic Programming. <http://mssanz.org.au>.

Kane, R.P., 1998. Extremes of the ENSO phenomenon and Indian summer monsoon rainfall. International Journal of Climatology 18, 775–791.

Koza, J.R., 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge, MA, USA.

Krishna Kumar, K., Rajagopalan, B., Cane, M.A., 1999. On the weakening relationship between the Indian Monsoon and ENSO. Science 284 (5423), 2156–2159. doi:10.1126/science.284.5423.2156.

Li, T., Zhang, Y.S., Chang, C.P., Wang, B., 2001. On the relationship between Indian Ocean sea surface temperature and Asian summer monsoon. Geophysical Research Letters 28, 2843–2846.

Liong, S.Y., Nguyen, V.T., Gautam, T.R., Wee, L., 2001. Alternative well calibrated rainfall–runoff model: genetic programming scheme. In: Paper Presented at World Water and Environmental Resources Congress 2001, Orlando, Florida, USA, pp. 777–787. doi:10.1061/40583(275)73.

Maity, R., Kashid, S.S., 2010. Short-term basin-scale streamflow forecasting using large-scale coupled atmospheric–oceanic circulation and local outgoing longwave radiation. Journal of Hydrometeorology 11 (2), 370–387.

Maity, R., Nagesh Kumar, D., 2006. Bayesian dynamic modeling for monthly Indian summer monsoon rainfall using ENSO and EQUINOO. Journal of Geophysical Research III, D07104. doi:10.1029/2005JD006539.

Maity, R., Nagesh Kumar, D., 2007. Hydroclimatic teleconnection between global sea surface temperature and rainfall over India at subdivisional monthly scale. Hydrological Processes 21, 1802–1813. doi:10.1002/hyp.6300.

Maity, R., Nagesh Kumar, D., 2008. Basin-scale streamflow forecasting using the information of large-scale atmospheric circulation phenomena. Hydrological Processes 22, 643–650. doi:10.1002/hyp.

Makkeasorn, A., Chang, N.B., Zhou, X., 2008. Short-term streamflow forecasting with global climate change implications – a comparative study between genetic programming and neural network models. Journal of Hydrology 352, 336–354.

Marcella, M.P., Eltahir, E.A.B., 2008. The hydroclimatology of Kuwait: explaining the variability of rainfall at seasonal and interannual time scales. Journal of Hydrometeorology 9, 1095–1105.

Minns, A.W., Hall, M.J., 1996. Artificial neural networks as rainfall–runoff models. Hydrological Sciences Journal 41 (3), 399–418.

Muttil, N., Liong, S.Y., 2001. Improving runoff forecasting by input variable selection in genetic programming. In: ASCE World Water Congress, vol. 111, Orlando, Florida, USA, 20–24 May, pp. 76–76. doi:10.1061/40569(2001)76.

Nageswara Rao, G., 1997. Interannual variation of monsoon rainfall in Godavari River basin – connections with the Southern Oscillation. Journal of Climate 11, 768–771.

Olivera, R., Loucks, D.P., 1997. Operating rules for multireservoir systems. Water Resources Research 33 (4), 839–852.

Ozelkan, E.C., Duckstein, L., 2001. Fuzzy conceptual rainfall–runoff models. Journal of Hydrology 253 (1–4), 41–68.

Parthasarathy, B., Diaz, H.F., Eischeid, J.K., 1988. Prediction of all India summer monsoon rainfall with regional and large-scale parameters. Journal of Geophysics Research 93 (5), 5341–5350.

Piechota, T.C., Dracup, J.A., Fovell, R.G., 1997. Western US streamflow and atmospheric circulation patterns during El Niño-Southern Oscillation. Journal of Hydrology 201, 249–271.

Raman, H., Sunilkumar, N., 1995. Multivariate modeling of water resources time series using artificial neural networks. Hydrological Sciences Journal 40 (2), 145–163.

Rasmusson, E.M., Carpenter, T.H., 1983. The relationship between eastern equatorial Pacific sea surface temperature and rainfall over India and Sri Lanka. Monthly Weather Review 111, 517–528.

Saji, N.H., Goswami, B.N., Vinayachandran, P.N., Yamagata, T., 1999. A dipole mode in the tropical Indian Ocean. Nature 401, 360–363.

Savic, D.A., Walters, G.A., Davidson, J.W., 1999. A genetic programming approach to rainfall–runoff modeling. Water Resources Management 13, 219–231.

Toth, E., Brath, A., Montanari, A., 2000. Comparison of short term rainfall prediction models for real time flood forecasting. Journal of Hydrology 239 (1–4), 132–147. doi:10.1016/s0022-1694(00)00344-9.

Wang, Q.J., 1991. The genetic algorithm and its application to calibrating conceptual rainfall–runoff models. Water Resources Research 27 (9), 2467–2471.

Wardlaw, R., Sharif, M., 1999. Evaluation of genetic algorithms for optimal reservoir system operation. Journal of Water Resources Planning and Management, ASCE 125 (1), 25–33.

Webster, P., Hoyos, C., 2004. Prediction of monsoon rainfall and river discharge on 15–30 day timescale. Bulletin of American Meteorological Society. doi:10.1175/BAMS-85-11-1745.

Whigham, P.A., Crapper, P.F., 2001. Modeling rainfall–runoff using genetic programming. Mathematical and Computer Modeling 33, 707–721 (Canberra, Australia).

Xie, P., Arkin, P.A., 1998. Global monthly precipitation estimates from satellite-observed outgoing longwave radiation. Journal of Climate 11, 137–164.

Xiong, L., Asaad, Y., Shamseldin, Y., O'Connor, K.M., 2001. A non-linear combination of the forecast of rainfall-runoff models by first order Takagi-Sugeno fuzzy system. Journal of Hydrology 254 (1–4), 196–217.

Yu, P.S., Chen, C.J., Chen, S.J., 2000. Application of gray and fuzzy methods for rainfall forecasting. Journal of Hydrologic Engineering, ASCE 5 (4), 339–345.