Short-Term Basin-Scale Streamflow Forecasting Using Large-Scale Coupled Atmospheric–Oceanic Circulation and Local Outgoing Longwave Radiation

RAJIB MAITY

Department of Civil Engineering, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal, India

S. S. KASHID

Department of Civil Engineering, Indian Institute of Technology Bombay, Powai, Mumbai, India

(Manuscript received 31 March 2009, in final form 24 August 2009)

ABSTRACT

This paper investigates the use of large-scale circulation patterns (El Niño–Southern Oscillation and the equatorial Indian Ocean Oscillation), local outgoing longwave radiation (OLR), and previous streamflow information for short-term (weekly) basin-scale streamflow forecasting. To model the complex relationship between these inputs and basin-scale streamflow, an artificial intelligence approach—genetic programming (GP)—has been employed. Research findings of this study indicate that the use of large-scale atmospheric circulation information and streamflow at previous time steps, along with OLR as a local meteorological input, potentially improves the performance of weekly basin-scale streamflow prediction. The genetic programming approach is found to capture the complex relationship between the weekly streamflow and various inputs. Different input variable combinations were explored to come up with the best one. The observed and predicted streamflows were found to correspond well with each other with a coefficient of determination of 0.653 (correlation coefficient $r = 0.808$), which may appear attractive for such a complex system.

1. Introduction

The management of land and water resources involves designing and operating water resources systems to cope with variability in rainfall and streamflow with time and space. Such variability in streamflow imposes many challenges in management of risks and opportunities associated with water resources systems. Reliable forecasts of streamflow a few weeks in advance can reasonably improve the management of water resources systems in rural as well as urban environments (Chiew et al. 2003). Operation of single purpose as well as multipurpose reservoirs largely depends upon the inflows into the reservoirs. The seasonal forecasts can also be useful for flood control reservoirs for deciding the storage and release schedules. However, the weekly forecasts of reservoir inflow can be of great help for judicious water allocations for various purposes like irrigation, hydropower, industry, and domestic use.

Corresponding author address: Dr. Rajib Maity, Assistant Professor, Department of Civil Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721302, West Bengal, India.
E-mail: rajib@civil.iitkgp.ac.in

DOI: 10.1175/2009JHM1171.1

a. Influence of large-scale circulation pattern

The association between large-scale climate circulation pattern and the hydrologic variables is termed ''hydroclimatic teleconnection'' (Chiew et al. 1998; Maity et al. 2007). It has been established in the recent years that the natural variation of hydrologic variables is linked with large-scale atmospheric circulation pattern 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 2008b). Effect of El Niño–Southern Oscillation (ENSO) on streamflow has been discussed by various researchers around the world. Dracup and Kahya (1994) developed a relationship between streamflow and La Niña events in the United States. Eltahir (1996) discussed El Niño and the natural variability of the flow of the Nile River in Egypt. Piechota et al. (1997) discussed western U.S. streamflow and atmospheric circulation patterns during ENSO. Chiew et al. (1998) discussed the effect of ENSO and Australian rainfall, streamflow, and drought. The effects of ENSO and Pacific interdecadal oscillation on the water supply in the Columbia River basin were studied by Barton and

Ramirez (2004). Chandimala and Zubair (2007), tried to predict streamflow and rainfall based on ENSO for water resources management in Sri Lanka. The effect of ENSO on streamflows has been also studied for 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 2006b). Douglas et al. (2001) attempted long-range forecasting of flows in Ganges based on ENSO information. Chowdhury and Ward (2004) studied the effect of ENSO on streamflows for the Greater Ganges–Brahmaputra–Meghna basins. Nageswara Rao (1998) studied interannual variation of monsoon rainfall in the 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 a 15–30-day time scale in the Brahmaputra and Ganges River basins. Maity and Nagesh Kumar (2008a) developed a scheme for basin-scale monthly streamflow forecasting by using the information of large-scale atmospheric circulation phenomena.

Another phenomenon, known as the Indian Ocean dipole (IOD) mode has been observed in the Indian Ocean, which is a pattern of internal variability with anomalously low sea surface temperatures off Sumatra and high sea surface temperatures over the western part of the tropical Indian Ocean (WEIO), with accompanying wind and precipitation anomalies (Saji et al. 1999; Webster et al. 1999). There are two coupled components of IOD—oceanic and atmospheric. The dipole mode index (DMI) is the oceanic component (Saji et al. 1999) of the IOD mode, which is defined as the difference in SST anomaly between the tropical western Indian Ocean (10°S–10°N, 50°–70°E) and the tropical southeastern Indian Ocean (10°S–0°, 90°–110°E).

The IOD mode has important applications in climate variability in the regions surrounding the Indian Ocean, like East Africa and Indonesia. Positive dipole event causes high rainfall over East Africa and drought in Indonesia and vice versa (Saji et al. 1999). However, statistical correlation of the DMI with precipitation over the Asian monsoon regime does not yield a significant relationship. Therefore, the relationship of the DMI to the Indian summer monsoon rainfall (ISMR) variability is not clear (Saji et al. 1999). However, Ashok et al. (2001) have shown that the IOD plays an important role as a modulator of the ENSO–ISMR relationship. In fact whenever the ENSO–ISMR correlation is low (high), the IOD–ISMR correlation is high (low) (Ashok et al. 2001).

Equatorial Indian Ocean Oscillation (EQUINOO) is the atmospheric component of the IOD mode (Gadgil et al. 2003, 2004). The anomalies in the sea level pressure and the zonal component of the surface wind along the equator are consistent with the convection anomalies. When the convection is enhanced (suppressed) over the western part of equatorial Indian Ocean (WEIO bounded by 10°S–10°N, 50°–70°E), it is suppressed (enhanced) over the eastern part of equatorial Indian Ocean (EEIO bounded by 10°S–0°, 90°–110°E) and the anomalous surface pressure gradient is toward the west (east) so that the anomalous surface wind along the equator becomes easterly (westerly). The oscillation between these two states is known as EQUINOO (Gadgil et al. 2004). Gadgil et al. (2004) have shown that the Indian summer monsoon rainfall is not only associated with ENSO, but also with EQUINOO. They suggest that the ISMR can be estimated by knowing the prior status of EQUINOO. 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 (2.5°S–2.5°N, 60°–90°E).

Nearly 80% of ISMR is due to the southwest monsoon in 4 months (June–September) and it is associated with various large-scale circulations over the oceans, which regulates the amount and distribution of the rainfall over the Indian subcontinent. However, such association is more prominently observed on a large geographical scale (continental and subcontinental) as compared to a small geographical scale like a river basin. Also, such an association is more prominently observed for a longer temporal scale (i.e., seasonal or monthly), as compared to a smaller temporal scale (i.e., 1–2 weeks). In other words, the strength of the hydroclimatic teleconnection decreases for a smaller spatiotemporal 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 2006b). The reasons behind the decreased strength of hydroclimatic teleconnection for a smaller spatiotemporal scale may include the local topography and weather systems prevailing in that region. 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. While there are certain common aspects of hydroclimatic teleconnection over the region, it may be modified because of the presence of such factors. 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.

While attempting weekly streamflow predictions based on large-scale atmospheric indices, it is not presumed that there must be exact week-to-week or day-to-day connection between large-scale indices over oceans and

basin-scale streamflow. However, there exists a systematic progressive teleconnection evolving with time, between sea surface temperatures, pressures, winds, and precipitation over the four months of the monsoon (Saji et al. 1999; Webster et al. 1999). It has been established in recent years that the natural variation of hydrologic variables is linked with these large-scale atmospheric circulation pattern through hydroclimatic teleconnection. Every year, the Indian monsoon is systematically developed on oceans and it progresses across a continent. In fact, this mechanism of progress of the monsoon is the key idea behind using lagged values of input variables in the analyses.

b. *Utility of outgoing longwave radiation*

Outgoing longwave radiation (OLR) is the energy leaving the earth as infrared (IR) radiation at low energy. OLR is strongly related to emission temperatures of radiating bodies at the surface and within the atmosphere. In clear sky this may be the surface, but more commonly OLR corresponds to the major cloud tops. This is especially true at low latitudes. Thus, OLR mostly measures cloud-top temperatures and only slightly less directly, given lapse rates that prevail at low latitudes, cloud-top heights. Deep clouds in these largely cumulus-convection-dominated regions correspond to more intense precipitation. Thus, OLR is used as a proxy for cumulus activity and precipitation in the region being studied and is used for this study. It is established that observed rainfall is best correlated with OLR at 10–30-day scales (Liebmann et al. 1998). The anomaly of OLR exhibits a negative correlation with precipitation over most of the globe (Xie and Arkin 1998). Rainfall estimates were reasonably consistent in both, distribution and magnitude over India, with climatological mean fields derived from rain gauge measurements (Arkin et al. 1989). Space–time variability analyses of the Indian monsoon rainfall as inferred from satellite-derived OLR data has been discussed by Haque and Lal (1991). Wang et al. (2002) have discussed the use of satellite observations for studying long-term changes in tropical clouds using OLR information from 1985 to 1998.

Summer rainfall in the tropics is usually associated with organized convective clouds, and these clouds modulate the OLR observed from satellite sensors. A procedure for obtaining rainfall rates from a mix of satellite- and surface-based observations has been outlined by Gairola and Krishnamurti (1992). They used OLR, microwave radiometric data, and surface rain gauge data for a composite analysis of rain rates.

Continuous monitoring of OLR anomalies from satellites over the global tropics offers an exciting potential for understanding large-scale rainfall variability. A large compilation of OLR data from the National Oceanic and Atmospheric Administration (NOAA), Earth Radiation Budget Satellite (ERBS), Nimbus, and Television and Infrared Observation Satellite-N (TIROS-N) [from the Advanced Very High Resolution Radiometer (AVHRR)] satellites are now available. Hence, regional studies for monitoring interannual variability in cloudiness and precipitation in relation to OLR can be useful in order to arrive at a predictive methodology in the realm of monsoon circulation.

In general, the Indian subcontinent has a high positive OLR anomaly during the winter and premonsoon months and a moderate negative anomaly in the monsoon season. The distribution of OLR anomalies exhibits strong dependence on both temporal and spatial scales during the years with poor or excess monsoon rainfall. Hence, this study includes the use of OLR as one of the important inputs for streamflow prediction.

c. *Objective of the paper*

The aim of this paper is to investigate the simultaneous influence of large-scale circulation pattern and local meteorological information on the weekly variation of basin-scale streamflow. The possible improvement of prediction performance of weekly streamflow using large-scale information, supported by local-scale meteorological information is thus investigated in this work.

In cases of ungauged catchments, nonrecording rain gauges, and in addition, the poor communication between the rain gauge stations and streamflow forecasting stations, accurate prediction of streamflow is a difficult task. Furthermore, in the recent past, researchers have indicated that the consideration of hydrometeorological and hydroclimatological input is helpful in streamflow forecasting (Makkeasorn et al. 2008). With the advent of various meteorological satellites orbiting around the earth, it is now possible to track the convection of cloud systems over the continents. Since monsoon cloud systems traversing over the river basin modulate the OLR observed from satellite sensors, the OLR anomaly can be used as a proxy to rainfall over the catchment for basin-scale streamflow prediction. Hence, the OLR information is used for streamflow prediction in this study, instead of any other form of direct input of rainfall information. However, streamflows at previous time steps were included as an input to the model, as serial correlations in successive streamflow values are always found to be significant.

Thus, weekly streamflow prediction using signals of large-scale atmospheric circulation patterns, local scale meteorological input (OLR), and streamflow at a few previous time steps has been attempted in this study.

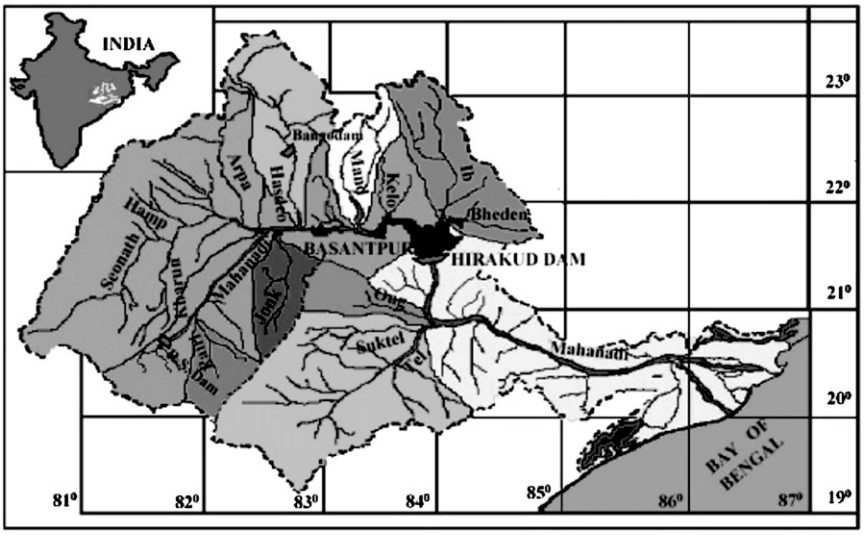

FIG. 1. Location map of the catchment and subbasins of the Mahanadi River (source: Central Water Commission, Orissa, India).

The historical average value of streamflow in a particular week is also provided as a long-term hydrological input.

2. Study area

The basin-scale streamflows referred to in this study are observed at the Basantpur River gauging site across Mahanadi River in India. The Mahanadi River rises in the highlands of Chhattisgarh and flows through Orissa to reach the Bay of Bengal. Hirakud is one of the most important multipurpose reservoirs in India, which is built across river Mahanadi. The location map of the Mahanadi catchment, the subbasins of the Mahanadi River and the Hirakud dam are shown in Fig. 1. The Hirakud dam intercepts 83 400 km^2 of Mahanadi catchment, which is about 65% of the total catchment area on Mahanadi. The river gauging station, Basantpur, is located just upstream of the Hirakud reservoir in the state of Orissa, India (Fig. 1). The stream gauging site is maintained by Central Water Commissions (CWC), Orissa, India. The catchment area up to this site consists of extensive areas of the Chattisgarh state of India.

3. Data

Sea surface temperature anomaly (SSTA) data from the Niño-3.4 region (5°S–5°N, 120°–170°W) are used as the ENSO index in this study. Weekly SSTA data is obtained from the Web site of the NOAA/National Weather Service Climate Prediction Center (see online at http://www.cpc.noaa.gov/data/indices/) for the period January 1990–December 2003. Similarly, EQWIN is used as EQUINOO index as explained earlier. Weekly surface wind data for the period January 1990–December 2003 is obtained from the National Centers for Environmental Prediction (NCEP; see online at http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.surface.html). Daily streamflow data at the Basantpur site is obtained from the office of Executive Engineer, Mahanadi Division, CWC, Orissa, India, for the period January 1990–December 2003.

The catchment area of the Mahanadi River at the Basantpur stream gauging site extends between 19° and 24°N latitudes and 80°–84°E longitudes. However, the OLR data used in this study for streamflow estimation in the Mahanadi basin were collected from the extensive region (15°–25°N, 75°–90°E) at 2.5° latitude–longitude intervals. The daily mean values of OLR over the Indian region, for 15 yr from 1 January 1990 to 31 December 2003 were used in this study. The extensions beyond the catchment were deliberately taken to capture the effect of advancing cloud systems across the basin over a period of time. The daily mean OLR data were used to derive weekly means. The long-term mean OLR was calculated first for the particular week. The weekly mean OLR for a specified region was obtained by summing the weekly grid values in a region and then dividing the sum by the number of grid points composing the region. The OLR anomalies for the region under consideration were then computed by deducting the average weekly OLR over the region from the observed OLR value for the particular week. Interpolated OLR data used in this study was obtained from the NOAA Web site (see online at http://www.cdc.noaa.gov).

4. Methodology

It is mathematically challenging to use climate signals for the prediction of basin-scale hydrologic variables because of 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, such as, Artificial Neural Networks (ANNs), (genetic algorithm) GA-based evolutionary optimizer, and genetic programming (GP). Thus, AI tools are tried nowadays for modeling complex systems. Application of ANN varies from 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. 2000), and regional annual runoff forecasting using indices of low-frequency climatic variability (Coulibaly et al. 2000). Ozelkan and Duckstein (2001) used the (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. The GA was also used in different fields of applications, for example, rainfall-runoff modeling (Wang 1991; Savic et al. 1999; Cheng et al. 2002), water quality models (Chau 2002), operation of multireservoirs systems (Olivera and Loucks 1997), and optimal reservoir operation (Wardlaw and Sharif 1999).

The GP is basically a 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 the GA (or rather a part of it) but unlike the latter, its solution is a computer program or an equation, as opposed to a set of numbers in the GA. Koza (1992) defines the 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.''

The search procedure of a model deals with a 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 the GP with 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 the ''black-box nature'' of the NN approach are issues of concern to many researchers. In an AR model, only the endogenous properties of the time series are used. To incorporate the external forcing, an option of an AR model with exogenous inputs (ARX) may be selected. However, if there are more exogenous inputs, computational complexity becomes 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. The 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 an ANN. The specialty of the 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), which is discussed later. However, GP also has some disadvantages. First, it is very computer intensive and requires extensive computing power. However, owing to the advent of fast computing facilities now available, this disadvantage can be handled. In GP, there are many ''possibly suitable'' programs. This may create a dubious attitude, as it seems to be difficult to select the best (single) program. However, if it is agreed upon that there could be many possible lines of attack to address a problem, having more than one ''possible program'' is really not an issue.

a. Genetic programming operators

The three general genetic operators are: crossover, reproduction, and mutation. Crossover and reproduction are responsible for the genetic diversity in the population of programs. Crossover is being operated on two programs, on which two random nodes are selected and the remaining part of the programs (subtrees) are swapped. The resulting two new programs are considered as a part of the next generation. Reproduction is performed by simply transferring a program from the old generation to the next generation without any change. Whereas the Darwinian reproduction operation creates a tendency toward convergence, the crossover operation exerts a counterbalancing pressure away from convergence in genetic programming. Thus, convergence of the population is unlikely in genetic programming (Koza 1992). Mutation is also an important operator in genetic algorithms to maintain diversity in the population. However, mutation may not be very important in the genetic programming (Koza 1992). This is because the dynamicity of individual programs (in

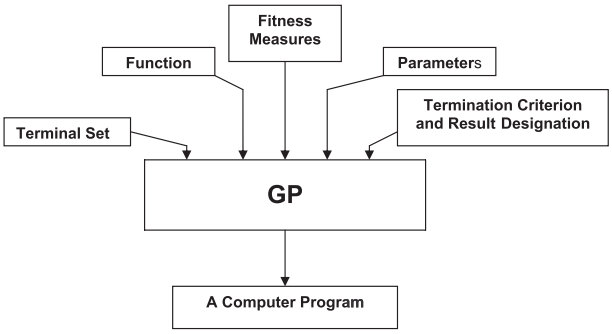

FIG. 2. Schematic diagram showing preparatory steps for the GP.

terms of sizes and shapes) already provides the required diversity in the population.

Selection of the individuals for crossover and reproduction can be done as per the genetic algorithm methodology of measuring fitness. This can be carried out in two different ways. First one can choose the individuals with the highest fitness for reproduction, which ensures the survival of the strong competitor. Another way is to assign a probability that a particular competitor will be selected for either reproduction or crossover. A more thorough discussion of these concepts is covered in Koza (1992).

b. Preparatory steps for application of the genetic programming

Application of the 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 a results and the criterion for terminating a run. A schematic diagram showing five major preparatory steps involved in GP is shown in Fig. 2.

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 following form:

$$Y^m = f(X^n), \tag{1}$$

where X^n is an n-dimensional input vector, and Y^m is an m-dimensional output vector. In the proposed study, the input and output variables are discussed later. The flowchart of genetic programming methodology is shown in Fig. 3 (Hong and Bhamidimarri 2003).

c. Program-based genetic programming

The results reported in the manuscript were based on the developed programs, specifically known as ''program-based GP'' (Frankone 1998). This approach is a data-driven approach, which develops a computer program to model the target output using the input variables during the training period. There are two types of computer programs written by the tool as a solution (viz. a ''program model'' and a ''team model''). A program model or an ''evolved program'' is a single program, which models the input data. A team model is a combination of few ''single program models,'' which are combined to produce a better result than any of the single program models. During the processing, the best programs are assembled into teams. The outputs from all of the programs that compose a team are assembled into one collective output that is frequently better than any particular member of the team. Most of the results presented in this analysis are the results of team models, which process the data rigorously to give the best possible results. This is the reason why the models developed by this methodology are better than traditional statistical modeling methodologies and other artificial intelligence tools, like artificial neural networks (Hong and Rosen 2002).

The ''impact frequency'' of every input variable is also identified in the program evolution process. After the complete analysis, the genetic programming looks through all of the programs in the team and analyzes how many times each input appears in a way that contributes to the fitness of the programs (Frankone 1998). The impact frequency is the percentage of times a variable is used in the best 30 programs evolved by the GP tool. For example, a value of impact frequency of 0.65, indicates that this input variable is used in 65% cases. Hence, even though the impact factor gives the importance of a particular variable in the evolution of the team of the best computer programs, it does not directly represent the exact coefficient or weighting to that input variable.

d. Genetic programming approach for weekly streamflow forecasting

A GP model is developed to predict weekly inflows at the Basantpur stream gauging site, across the Mahanadi River in the state of Orissa in India. A total of 11 analyses were carried out to find out the most influential input variables and the best model for the prediction of weekly streamflow of the Mahanadi River. The analyses were based on information of large-scale atmospheric circulations in the form of weekly values of ENSO indices, EQUINOO indices, and OLR anomalies for certain number of previous time steps, which will be discussed later.

Significant serial correlation exists between the streamflow of the current time step and the the streamflow of previous time steps during June–October. Hence, streamflows of previous time steps are also used. The

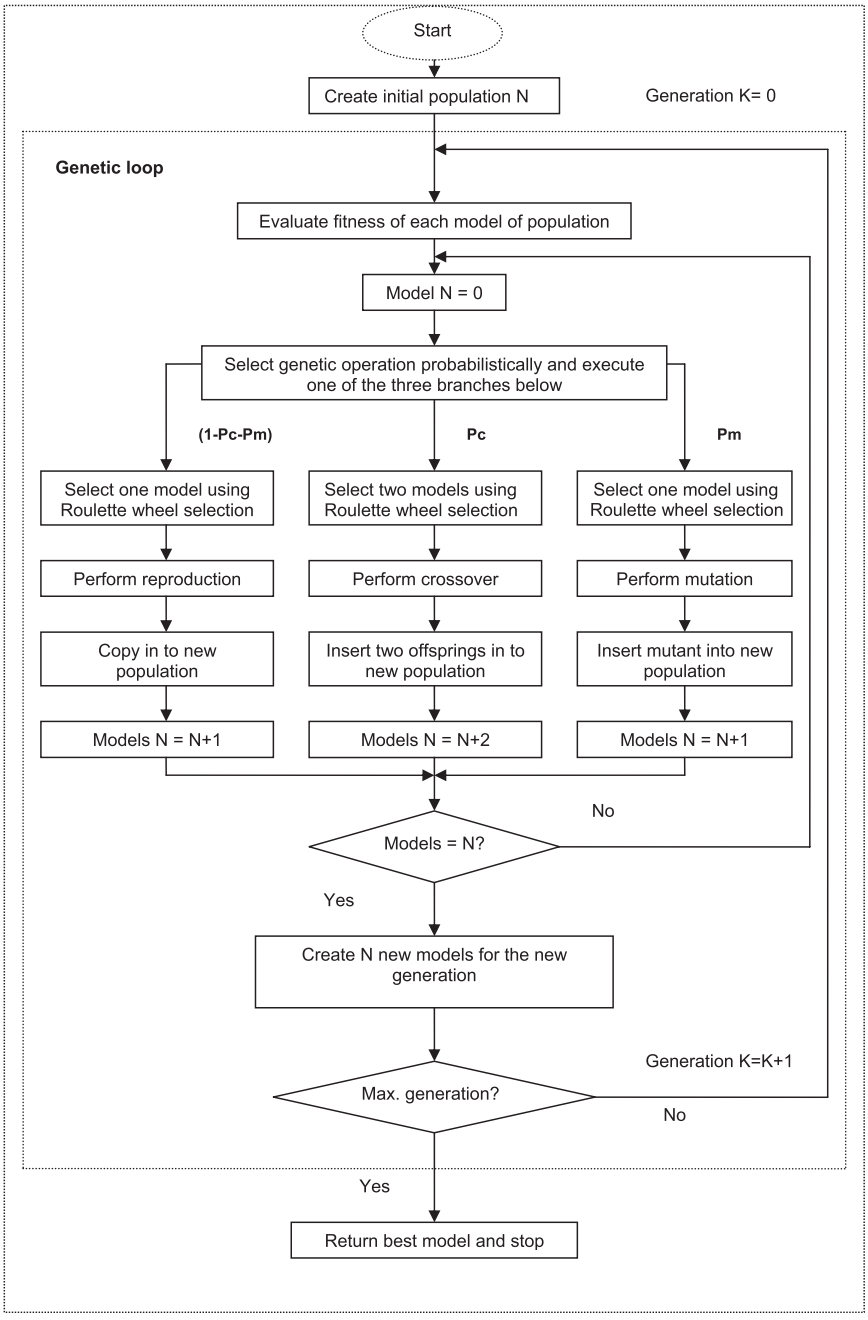

FIG. 3. Flowchart for genetic programming (Hong and Bhamidimarri 2003).

historical average of the streamflow at the current time step is also used. The weekly data is designated with respect to the central date of that week. For example, the week shown as 28 June 1995 indicates the week runs from 25 June 1995 to 1 July 1995. The immediate next week will be designated as 5 July 1995.

Weekly data from 1 January 1990 to 31 December 2003 were used for this study. Daily streamflow data were available since January 1972 at the Basantpur site, but the daily OLR data were only available since January 1990. Hence, the period of analysis had to be limited to 1990–2003. From this period, weekly data of 162 monsoon weeks of 1990–98 were used for training purposes and data of 90 monsoon weeks of 1999–2003 were used for testing purposes. Overfitting in GP analysis is observed if there is relatively less training datasets compared to the number of inputs. In this case, 252 sets of weekly data (14 yr $\times$ 18 monsoon weeks yr^{-1}) have

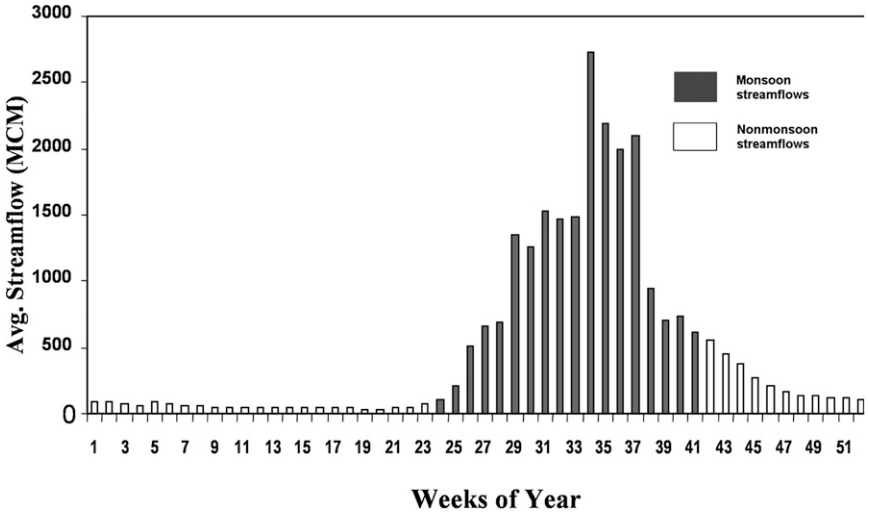

FIG. 4. Historical weekly average streamflow.

been used, which is a good quantum of data for using GP, without any fear of overfitting.

The historical weekly average streamflow in the Mahanadi River at the Basantpur site is shown in Fig. 4, where MCM stands for millions of cubic meters (10^6 m^3). It can be observed from Fig. 4, that, on an average, monsoon streamflows were significant from the third week of June (24th week in Fig. 4). Monsoon streamflows for 18 successive weeks, starting from the third week of June, are considered for weekly streamflow prediction.

Generally groundwater contribution to the streamflow for a large watershed is significant and needs to be considered separately. The watershed considered in this study is a rain-fed basin and the contribution of the groundwater is almost nil. This is reflected by the nonmonsoonal flow pattern (Fig. 4) also reported in an earlier project report of the CWC (Patri 1993). However, the subsurface flow is significant for the monsoon months and is incorporated by the streamflow information from the previous step(s). This is because prevailing conditions–characteristics of the watershed are important factors for the future streamflow status, apart from all other external forcing mechanisms. The prevailing conditions–status may include antecedent water content, subsurface flow, groundwater contribution, etc. Streamflow information from previous time steps can be considered as an indirect measure of these factors, which is generally considered by traditional modeling approaches. However, use of only this information may not provide satisfactory results. The use of exogenous inputs may provide significant improvement, which is the focus of this study and discussed in detail later.

The weekly streamflow is modeled as a function of (i) the historical average weekly streamflow for the particular week, (ii) the streamflow of certain number of previous weekly time steps, (iii) the ENSO index of a certain number of previous weekly time steps, (iv) the EQUINOO index of a certain number of previous weekly time steps, and (v) the OLR anomaly over the basin for certain number of previous weekly time steps. Thus, an expanded form of Eq. (1), can be written as

$$\mathrm{SF}_t = f\begin{bmatrix}\mathrm{HSF}_t, (\mathrm{SF}_{t-1}, \mathrm{SF}_{t-2}, \cdots), (\mathrm{EN}_{t-1}, \mathrm{EN}_{t-2}, \cdots),\\ (\mathrm{EQ}_{t-1}, \mathrm{EQ}_{t-2}, \cdots), (\mathrm{OLR}_{t-1}, \mathrm{OLR}_{t-2}, \cdots)\end{bmatrix}, \quad (2)$$

where SF stands for streamflow, HSF stands for the historical weekly average streamflow, EN stands for the ENSO index, EQ stands for the EQUINOO index, and OLR stands for the OLR anomaly. The optimum number of lags to be considered for each input variables is decided based on the ''input impacts'' of that input variable during model calibration as discussed later. It should be noted here that several weeks of lags are used to incorporate evolutionary trends in the data for all the input variables except HSF (which is of course a single value for a particular week). The reason behind considering the lagged variables is discussed in the next section.

5. Results and discussion

The phenomenon of a monsoon causing rainfall over the Indian subcontinent evolves in the Pacific Ocean and Indian Ocean. The evolution period extends back even more than 3 months before the ongoing monsoon month (June–September). The strength of the progressing monsoon is hence captured in terms of ENSO and EQUINOO indices at previous time steps in real time with the

TABLE 1. Different combinations of input variables for streamflow prediction: historical monthly average streamflow (HASF), outgoing longwave radiation (OLR), streamflow (SF), ENSO index (EN), EQUINOO index (EQ). The mark √ indicates variable is included. Subscripts to the variables indicate time lag in which *t* corresponds to the current time.

		Variable combinations										
Variable No.	Variable	C-1	C-2	C-3	C-4	C-5	C-6	C-7	C-8	C-9	C-10	C-11
1	HSF	√		√	√	√	√	√		√	√	√
2	OLR_{t-2}	√			√	√	√		√	√		√
3	OLR_{t-1}	√			√	√	√	√	√	√		√
4	SF_{t-10}	√										
5	SF_{t-9}	√										
6	SF_{t-8}	√										
7	SF_{t-7}	√										
8	SF_{t-6}	√										
9	SF_{t-5}	√										
10	SF_{t-4}	√										
11	SF_{t-3}	√			√				√		√	√
12	SF_{t-2}	√			√				√		√	√
13	SF_{t-1}	√		√	√		√	√	√		√	√
14	EN_{t-12}	√										
15	EN_{t-11}	√										
16	EN_{t-10}	√				√						
17	EN_{t-9}	√				√						
18	EN_{t-8}	√				√						
19	EN_{t-7}	√	√	√		√	√	√	√		√	√
20	EN_{t-6}	√	√	√		√	√	√	√		√	√
21	EN_{t-5}	√	√	√		√	√	√	√		√	√
22	EN_{t-4}	√	√	√		√	√	√	√		√	√
23	EN_{t-3}	√	√	√		√	√	√	√		√	√
24	EN_{t-2}	√	√	√		√	√	√	√		√	√
25	EN_{t-1}	√										
26	EQ_{t-12}	√										
27	EQ_{t-11}	√										
28	EQ_{t-10}	√										
29	EQ_{t-9}	√										
30	EQ_{t-8}	√										
31	EQ_{t-7}	√	√	√			√	√	√		√	√
32	EQ_{t-6}	√	√	√			√	√	√		√	√
33	EQ_{t-5}	√	√	√			√	√	√		√	√
34	EQ_{t-4}	√	√	√		√	√	√	√		√	√
35	EQ_{t-3}	√	√	√		√	√	√	√		√	√
36	EQ_{t-2}	√	√	√		√	√	√	√		√	√
37	EQ_{t-1}	√				√						

progressing monsoon. Similarly, it can be understood that precipitated water over the catchment needs a few weeks to pass the stream gauging station. Also, OLR lags are useful for considering the increasing (decreasing) trend of precipitation in the previous weeks. Hence, OLR provides a few time steps for the previous (weekly) precipitation for streamflow forecasting. Similarly, lag streamflows are an indication of the trend of the streamflow in prior weeks. HSF for a particular week is an input to the model that incorporates the climatological value of the streamflow for that particular week.

Various combinations of input variables are carried out (Table 1) for evolving the best combination of input variables to end up with the best possible results using minimum numbers of input. To achieve this, the ability of the GP to indicate the significance of every input variable from all the given inputs to the program is used. The GP can evaluate the impact of each input variable that in turn helps to judge their importance and its further inclusion or exclusion in the next combination is decided. Thus, it is used as a screening tool for proposed input variables.

However, for the starting combination C-1, the input variables are used with sufficiently long lags for each input variable. The variables with better impact frequencies as well as a priori knowledge of casual variables and physical insight into the problem, different sets of input combinations, are selected for the successive analyses, which showed systematic stepwise improvements

TABLE 2. Impact frequencies of different input variables for different combinations: historical monthly average streamflow (HASF), outgoing longwave radiation (OLR), streamflow (SF), ENSO index (EN), EQUINOO index (EQ). Results for the best combination are shown in boldface. Subscripts to the variables indicates time lag, in which t corresponds to the current time.

		Input variable combinations										
Variable No.	Variable	C-1	C-2	C-3	C-4	C-5	C-6	C-7	C-8	C-9	C-10	C-11
1	HSF	1.00	1.00	0.80	1.00	1.00	0.83	0.97		1.00	0.63	**1.00**
2	OLR_{t-2}	0.13			0.63	0.67	0.43		0.77	1.00		**0.57**
3	OLR_{t-1}	0.50			0.97	0.60	0.67	0.90	0.97	1.00		**0.53**
4	SF_{t-10}	0.20										
5	SF_{t-9}	0.83										
6	SF_{t-8}	0.40										
7	SF_{t-7}	0.70										
8	SF_{t-6}	0.43										
9	SF_{t-5}	0.30										
10	SF_{t-4}	0.43										
11	SF_{t-3}	1.00			1.00				0.40		0.73	**0.37**
12	SF_{t-2}	0.20			0.33				0.17		0.30	**0.23**
13	SF_{t-1}	1.00		1.00	1.00		1.00	1.00	1.00		1.0	**1.00**
14	EN_{t-12}	0.23										
15	EN_{t-11}	0.00										
16	EN_{t-10}	0.23				0.50						
17	EN_{t-9}	0.17				0.67						
18	EN_{t-8}	0.03				0.60						
19	EN_{t-7}	0.33	0.00	0.10		0.37	0.40	0.27	0.33		0.13	**0.33**
20	EN_{t-6}	0.37	0.53	0.37		0.90	0.43	0.53	0.43		0.53	**0.60**
21	EN_{t-5}	0.47	0.47	0.53		1.00	0.77	0.57	0.40		0.40	**0.47**
22	EN_{t-4}	0.27	0.43	0.10		0.73	0.43	0.37	0.40		0.33	**0.43**
23	EN_{t-3}	0.27	0.20	0.47		0.60	0.43	0.73	0.57		0.30	**0.30**
24	EN_{t-2}	0.27	0.43	0.63		0.60	0.67	0.37	0.17		0.33	**0.53**
25	EN_{t-1}	0.10										
26	EQ_{t-12}	0.27										
27	EQ_{t-11}	0.57										
28	EQ_{t-10}	0.23										
29	EQ_{t-9}	0.30										
30	EQ_{t-8}	0.47										
31	EQ_{t-7}	0.10	0.60	0.53			0.37	0.60	0.30		0.30	**0.50**
32	EQ_{t-6}	0.67	0.57	0.63			0.40	0.87	0.57		0.50	**0.40**
33	EQ_{t-5}	0.27	0.63	0.43			0.43	0.47	0.33		0.53	**0.30**
34	EQ_{t-4}	0.60	0.43	0.67		0.70	0.97	0.73	0.73		0.80	**0.90**
35	EQ_{t-3}	0.10	0.23	0.27		0.97	0.53	0.93	0.57		0.30	**0.60**
36	EQ_{t-2}	0.60	0.50	0.60		1.00	0.53	0.60	0.43		0.17	**0.47**
37	EQ_{t-1}	0.33				0.57						

in the performance of the models. The results are compared by computing the aforementioned statistical measures during testing periods. The comparative statement in terms of impact frequencies of different variables in different combinations of variables are given in Table 2. The error statistics computed for different variable combinations is given in Table 3. The design of each combination and their results are discussed in the following subsections.

It is worthwhile to mention here that, in India, more than 80% of the annual rainfall is received in a limited period of 4 months of the monsoon season (June–September). Streamflow analysis is restricted to monsoon streamflows, as our major interest is during the monsoon period. And there is hardly any or no rainfall anywhere on the basin during the nonmonsoon period. The climatic systems causing the rainfall are roughly similar over the monsoon season. Thus, the models are assumed to be of homogeneous type.

It is also worthwhile to mention here, that the inherent noise of the inputs is not considered in such a modeling approach. The uncertainty quantification for the predicted values, including uncertainty arising from noise in input data, is not in the scope of the paper. However, it can be mentioned that the presence of noise in the input data will increase the uncertainty associated with the predicted values as it is additive with the other sources of uncertainty (e.g., model uncertainty). It is not possible to quantify the

TABLE 3. Error statistics for different combinations of input variables: coefficient of determination (r^2), mean absolute error (MAE; in units of MCM), root-mean-squared error (RMSE; in MCM), training period (train), and testing period (test).

		Input variable combinations										
	Statistical Measures	C-1	C-2	C-3	C-4	C-5	C-6	C-7	C-8	C-9	C-10	C-11
1	r^2 (Train)	0.833	0.655	0.743	0.595	0.740	0.587	0.810	0.621	0.555	0.722	0.695
2	r^2 (Test)	0.567	0.589	0.596	0.529	0.433	0.592	0.621	0.565	0.367	0.613	0.653
3	MAE (train)	404	546	494	572	511	565	439	558	588	454	501
4	MAE (test)	498	533	488	481	595	469	470	502	644	450	432
5	RMSE (train)	577	783	696	822	707	855	606	799	895	714	755
6	RMSE (test)	723	720	702	755	825	716	689	726	914	691	655

associated uncertainty with deterministic models and the proposed model is deterministic in nature. Use of input data after the removal or reduction of noise is expected to have a positive effect on overall prediction performance. In this study, whatever data is available from the data source (mentioned in section 3) is used.

a. Gradual evolution of different combinations

1) INPUT VARIABLE COMBINATION C-1

The very first combination of input variables is used with a ''sufficient'' number of lags with a goal to achieve the optimum number of lags in the subsequent combinations. Hence, the optimum number of lags could be selected to obtain the best results in the model development process. The first combination C-1 for streamflow forecasting, thus, includes ENSO and EQUINOO over the last 12 weeks, streamflows over the last 12 weeks, and OLR anomalies over the last 2 weeks. The plots between the observed and predicted weekly streamflow, during the training and testing periods are shown in Figs. 5a,b, respectively. These plots will be compared with the final combination and a related discussion will be presented.

It was observed that every input had a different impact frequency. The impact frequencies for streamflow beyond three previous time steps, for the ENSO index beyond seven previous time steps, and for EQUINOO index indices beyond the seven weekly time steps were found to be considerably diminishing. Hence, ENSO index for time steps $t - 7$ to $t - 1$, EQUINOO index for time steps $t - 7$ to $t - 1$, and OLR anomalies for $t - 2$ to $t - 1$ were found to be the most significant inputs for streamflow prediction (Table 2). Depending upon the computed input impacts, the significant variables were selected for other combinations. It is also observed from Table 2, that the impact frequency of SF_{t-7} and SF_{t-9} from previous streamflow information and EQ_{t-8} and EQ_{t-11} from EQUINOO indices are also considerable. However, it may be noted that the genetic programming tool calculates the ''input impacts,'' based on the number of times the particular input variable was included in ''best 30 GP evolved programs,'' giving the best results. Though it is logical to use this feature of the GP to identify the most influential variables, the selection of input variables should also be supported by the priori knowledge of casual variables and physical insight into the problem. For instance, the previous studies indicates that the effect of EQUINOO is more immediate than the effect of ENSO on the Indian hydrologic phenomena. This is also convincing in the perspective of the geological locations (Maity and Nagesh Kumar 2006b). As the effect of ENSO is considered up to 7 lags (approximately 2 months), the EQUINOO is not considered beyond lag 7, even though for some lags (beyond 7) impact frequency is considerable.

Similarly, the streamflow information is used up to three previous time steps. The reason being the distance of gauging site about 500 km from the origin of the Mahanadi River. A substantial part of the generated streamflow passes the gauging site in 2–3 weeks. Also the influence of the previous steps of the streamflow is expected to be contiguous in time, unless there is some existence of special physiographic characteristics of the watershed, which is not apparent in the case of the Mahanadi River. Thus, even though input impact frequencies were considerable for inputs SF_{t-7} and SF_{t-9}, both the inputs were dropped. Still, prediction performances are also investigated if these lags are considered while considering streamflow information [please refer to section 5a(4) for details] even though the rationality of this combination is not supported by the physical insight.

2) INPUT VARIABLE COMBINATION C-2

Input variable combination C-2 considers just the ENSO and EQUINOO indices over a period of $t - 7$ to $t - 2$ time steps to investigate the usefulness of other inputs along with these large-scale circulation patterns. It is found that 58.9% ($r^2 = 0.589$) of the variability in the weekly streamflow can be explained (Table 3). This may indicate that the large-scale patterns have some influence

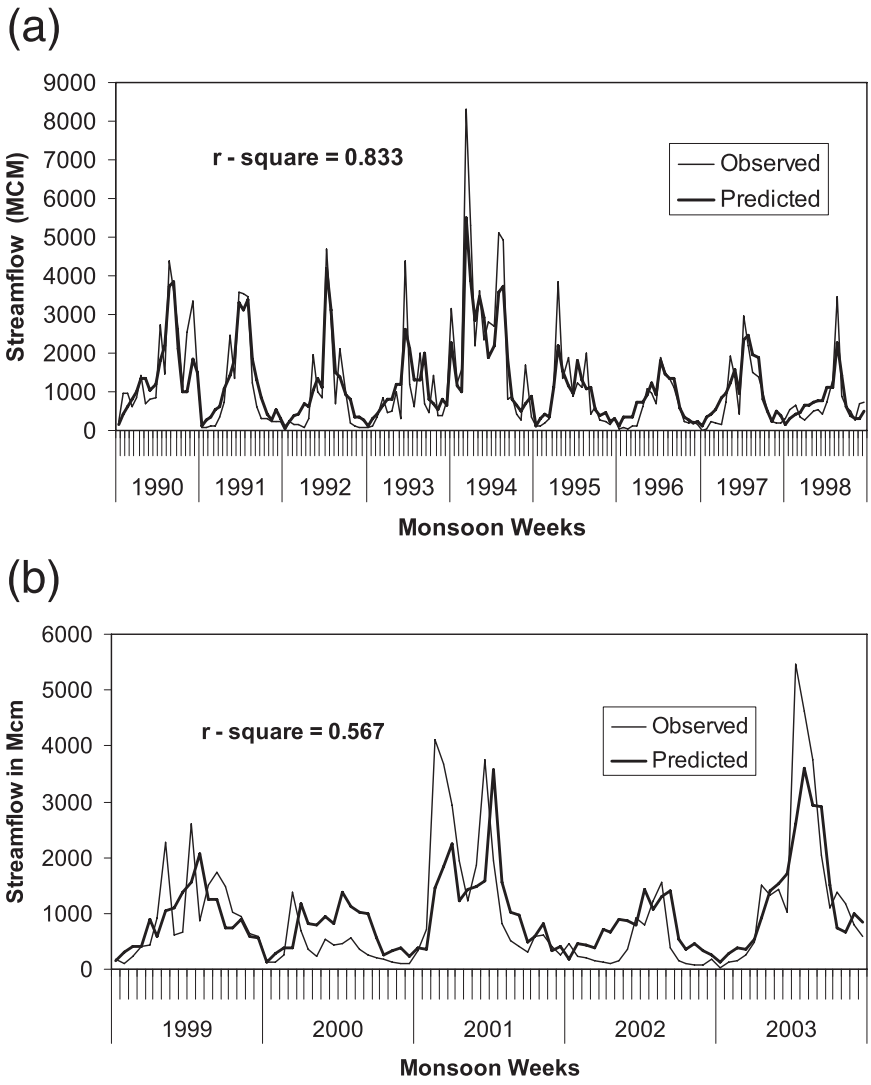

FIG. 5. A comparison between the observed and predicted streamflow during (a) the training period and (b) the testing period for the input variable combination C-1.

on the basin-scale hydrology. However, whether this influence is suppressed by the local influences or not is yet to be inferred. In the following combinations, it is investigated whether the large-scale atmospheric circulation indices are to be supported by some additional inputs for the maximum possible predictions performance.

3) INPUT VARIABLE COMBINATION C-3

Input variable combination C-3 includes the streamflow at the previous weekly single time step $t - 1$, in addition to inputs of combination C-2. A small improvement in the performance is observed being the r^2 value improved from 0.589 to 0.596 in comparison with combination C-2 (Table 3). Though the improvement is small, it indicates that the inclusion of the streamflow of a few more previous time steps may be useful, which will be investigated later. Before that a combination that excludes the large-scale information is carried out.

4) INPUT VARIABLE COMBINATION C-4

Input variable combination C-4 aims to investigate the performance while excluding the large-scale circulation information, ENSO and EQUINOO. Thus, the analysis is carried out for the performance of streamflow prediction, based on the historical streamflow (HSF), the streamflow information of previous time steps (SF_{t-1} to SF_{t-3}), and OLR (OLR_{t-1} and OLR_{t-2}) only. It is observed that prediction performances were poorer in training as well as testing phases because only 52.9% of the variability is explained ($r^2 = 0.529$) as shown in Table 3. The peak streamflows of 1994 and 1998 in training and 2001 and 2003 in testing were largely

underestimated (figure not shown). As discussed before, streamflow information, SF_{t-7} and SF_{t-9}, are not considered even though input impact frequencies were considerable in C-1. However, even though these lags are considered (discarding the rationality of the combination is not supported by physical insight, i.e., the considering the input set as C-4 plus SF_{t-7} and SF_{t-9}, the prediction performance was found to be 65.5% during training and 59.6% during the testing period in terms of explained variability). Even though the performances are better than the existing C-4, its performance is still not satisfactory (please refer to the full model later). It underlines the importance of using large-scale atmospheric circulation information in streamflow estimation models. This also indicates that it is necessary to use the large-scale atmospheric circulation indices along with the previous time step streamflow and local meteorological information for better predictions.

5) INPUT VARIABLE COMBINATION C-5

The combination C-5 tests the usefulness of the lags, based upon the input impacts compared with an earlier combination of C-3. The lags for ENSO and EQUINOO are changed to observe its effect on the performance and streamflow information from previous time steps are ignored despite their high impact factor. For ENSO, the lags are considered from $t - 10$ to $t - 2$ time steps (i.e., more information is considered) and for EQUINOO, the same is selected from $t - 4$ to $t - 1$ time steps. It is observed that prediction performance is very poor as compared to the earlier combination C-3 as only 43.3% of the variability is explained ($r^2 = 0.433$) as shown in Table 3. Thus, the input impacts provide useful information to select the proper lags for the input variables.

6) INPUT VARIABLE COMBINATION C-6

Based on the previous results in combinations C-2 and C-3, ENSO and EQUINOO indices from $t - 7$ to $t - 2$, streamflow information of one previous time step, and OLR for $t - 2$ to $t - 1$ time steps are used in combination C-6. It is observed that there is a substantial improvement in the prediction performance as compared to combination C-4, where large-scale circulation information—ENSO and EQUINOO—were excluded. The variability explained is found to be 0.592 during the testing period (Table 3). This indicates that the large-scale atmospheric circulation information is also important for the basin-scale hydrology.

7) INPUT VARIABLE COMBINATION C-7

The combination C-7 is an attempt to check whether OLR information from a single lag is sufficient or not. Thus, OLR information is considered only from the previous week (i.e., $t - 1$) and all other inputs are kept the same. There was marginal improvement in the performance. The value of input impact for OLR for the $t - 1$ step showed considerable improvement indicating the impact of OLR $t - 1$ to be a more important factor as compared to OLR $t - 2$ in the analyses. However, considering the physical reasoning that the rainfall in the last two weeks may be useful, OLR information from both the weeks are considered in the next combination.

8) INPUT VARIABLE COMBINATION C-8

A special variable combination is used to check the applicability of the model in the absence of a reasonable long dataset to calculate the historical streamflow values. Thus, historical average streamflow (climatological value) is excluded from the input variable combination. Thus, the input variable combination for C-8 consists of the streamflow information of three previous time steps, the ENSO and EQUINOO indices for $t - 7$ to $t - 2$ time steps and OLR for $t - 2$ to $t - 1$ time steps. It is observed that 56.5% ($r^2 = 0.565$) of the variability is explained by considering these inputs (Table 3). Thus, the exclusion of historical average streamflow ends up in a poorer performance, however, not completely useless. This indicates that the model can still be used in the absence of a sufficiently long historical record of the streamflow values, which might be a very common case in developing countries.

9) INPUT VARIABLE COMBINATION C-9

This input variable combination is tried to check whether a combination of only local meteorological variable OLR and historical average streamflow is able to provide reasonably good results. Thus, it is a performance trial in absence of the ENSO and EQUINOO indices and streamflow at previous time steps. It is observed that just 36.7% ($r^2 = 0.367$) of the variability is explained by this combination (Table 3). This result indicates that the local meteorological information alone cannot be sufficient for the prediction of basin-scale streamflow. This might be because prevailing conditions–characteristics of the watershed are also important factors for streamflow prediction. Previous time steps in streamflow information can be considered as an indirect measure of basin storage as well as groundwater contribution to the streamflow at current time steps. Moreover, large-scale external forcing from the remote atmospheric–oceanic circulation information through hydroclimatic teleconnection is also important. Thus, the simultaneous use of large-scale as well as local information including the previous time step streamflow may be necessary to provide the better predictions.

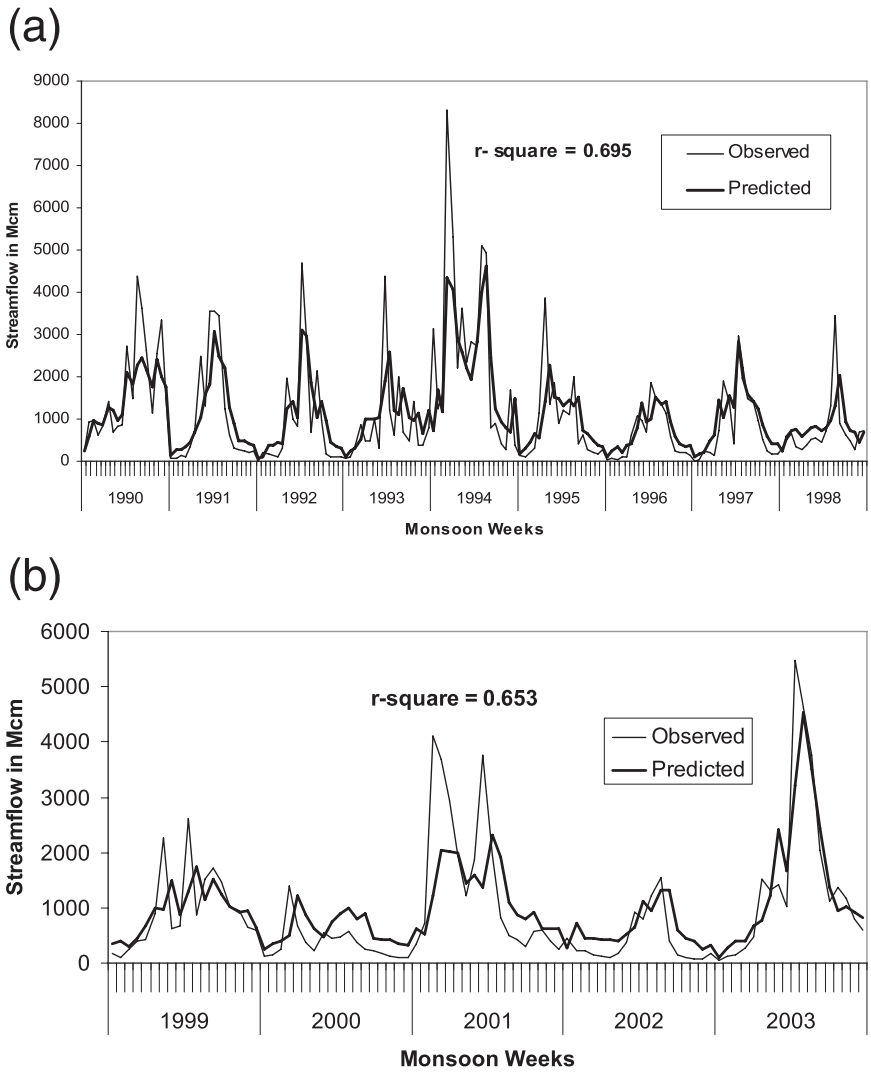

FIG. 6. As in Fig. 5, but for combination C-11.

10) INPUT VARIABLE COMBINATION C-10

After deciding on the optimum number of time steps for local as well as large-scale climate inputs, the next combinations are designed to estimate the contribution of OLR in basin-scale streamflow prediction. The input variable combination for C-10 includes streamflow information of the three previous time steps, the ENSO and EQUINOO indices for $t - 7$ to $t - 2$ time steps, but OLR information is excluded. It is observed that 61.3% ($r^2 = 0.613$) of the variability is explained by this input combination (Table 3), which appears to be the best combination so far. However, additional inclusion of local meteorological information in terms of OLR will be interesting, which is carried out in the next combination.

11) INPUT VARIABLE COMBINATION C-11

Input variable combination C-11 includes OLR information along with other input variables in C-10. It is observed that 65.3% ($r^2 = 0.653$) of the variability is explained by considering these inputs (Table 3). A comparison between C-10 and C-11 indicates that additional inclusion of OLR improves the explained variability to 65.3% from 61.3%. Thus, the improvement can be attributed to the inclusion of local meteorological information in terms of OLR, over the other large-scale inputs. The plots between the observed and predicted weekly streamflow, during the training and testing periods are shown in Figs. 6a,b, respectively. For comparison, the results of the initial full set of C-1 are shown in Fig. 5a for the training period and Fig. 5b for the

testing period. A comparison between these figures show that the extreme values are better captured in C-11 during the model testing period. This is an interesting point that even though all the inputs are considered in C-1, streamflow anomalies are captured with more accuracy in the case of C-11.

It may also worthwhile to check, whether the inclusion of historical average weekly streamflow for a particular week (climatological value) as an input variable, adds some artificial skill to the model. As indicated in the results of combination C-11, 65.3% variability is explained by this input variable combination. Now the explained variance between historical weekly average streamflow values and observed weekly streamflow values is also calculated. The coefficient of determination was found to be 0.1675. This indicates that only 16.75% of variance is explained by the climatological values, and the remaining 48.55% of variance is explained by the full model. This is due to the fact that at weekly scale, the variation is quite high. This is also the reason behind the difficulties in prediction by the traditional approach.

The deviations of weekly streamflow from climatology are also analyzed. The correlation coefficient "r" between anomalous values of streamflow (i.e., observed anomaly and predicted anomaly) is also calculated to check the robustness of the model. The correlation coefficient between the observed anomaly and the predicted anomaly is found to be 0.7863 ($r^2 = 0.618$). This indicates that the deviations of weekly streamflow from climatology are captured with reasonable accuracy. Because the inherent complexity is very high, as indicated earlier, the achieved accuracy of the predicted anomaly can be appreciable. This demonstrates the ability of the GP tool to capture the information from various inputs and provide reasonably reliable predictions.

In addition to the aforementioned combinations, three separate input variable combinations were also carried out (not reported in Table 3) to investigate the performance if only a single input is considered at a time. The combinations are (i) HSF and SF ($t - 3$ to $t - 1$), (ii) HSF and EN ($t - 7$ to $t - 2$), as well as (iii) HSF and EQ ($t - 7$ to $t - 2$). For these input combinations, the explained variability is found to be 46.7%, 39.1%, and 30.0%, respectively. In the case of the combination using HSF and OLR ($t - 2$ to $t - 1$; i.e., reported earlier as C-7), the explained variability was 38.9%. The results of these four special combinations can be used as an indication of their individual importance of the inputs even though it is not exact quantification. The results indicate the highest contribution of previous time step streamflow information toward the basin-scale streamflow prediction. However, the importance of local information in terms of OLR is also comparable to that from large-scale inputs.

b. *Major observations from gradually evolving combinations*

From the above analyses and discussions, the major point of the concurrent effect of large-scale and local influences on basin-scale streamflow can be drawn. Traditional modeling techniques generally use the historical record for the future prediction. However, the use of exogenous inputs provides significant improvement, if some cause–effect relationship is established. Such causal variables may include both local as well as large-scale climatic information. The ENSO and EQUINOO are used as large-scale influences and the OLR is used as local meteorological information. However, while considering the information of external forcing alone, the prediction performance was not the best one. This is due to the fact that prevailing conditions–characteristics of watershed are also important factors for streamflow prediction and such factors should be considered. Previous time step streamflow information can be considered as an indirect measure of basin storage as well as groundwater contribution to the streamflow at the current time steps. Observations from different evolving combinations indicate that the best prediction is achieved while also considering the previous time step streamflow information. Thus, current streamflow is dependent on antecedent water storage as well as external forcing. This concurrent influence might be one of the reasons behind the long-recognized complexity and nonlinearity in the soil–water and streamflow process. The proposed GA-based approach captures the complexity, at least to some reasonable extent, and establishes that the simultaneous use of the previous time step streamflows and local as well as large-scale information provides better forecasts.

The use of the historical average streamflow value ensures the use of existing seasonality in the streamflow values. The fluctuation over the seasonal values is supposed to be affected by some of the local and large-scale influences. However, as mentioned before, use of only streamflow information may not provide the satisfactory results. For example, as mentioned before in combination C-4, the peak streamflows of 1994 and 1998 in training and 2001 and 2003 in testing were largely underestimated, even the OLR was considered along with the previous time step streamflow information. It underlines the importance of using large-scale atmospheric circulation information in streamflow estimation models.

In combination C-2, historical average streamflow, ENSO (EN_{t-7} to EN_{t-2}), and EQUINOO (EQ_{t-7} to EQ_{t-2}), are considered and the model performance was found

to be 0.589 in terms of the coefficient of determination during the testing period (Table 3). On the other hand, while considering only local variables [i.e., the historical average streamflow, OLR (OLR_{t-2} and OLR_{t-1}), and previous time step streamflows (SF_{t-3}, SF_{t-2}, and SF_{t-1})] in combination C-4, the model performance was found to be 0.529 in terms of coefficient of determination during the testing period (Table 3). However, the performances of both these sets are inferior to the final combination C-11, in which a combination of all possible influencing factors is considered simultaneously. For example, the OLR provides the convective activity over the basin, the previous time step streamflow provides a proxy of the antecedent water content of the watershed, and the large-scale information provides the information of global forcing on local hydrologic processes. However, their contribution toward the predictability of the next time step streamflow need not necessarily be the same.

The model performance for the best combination (combination C-11), is shown in Fig. 6b during the testing period. It is found that the peak values are underestimated for some weeks in 1994 and 2001, which can be attributed to some other meteorological disturbances. However, the peak values in 2003 are well captured. Overall, the observed and predicted streamflows are well associated and a coefficient of determination was found to be 0.653 (a correlation coefficient of 0.808), which is reasonably appreciable for such a complex system.

6. Conclusions

Established research indicates an association between the large-scale circulation pattern and hydrologic variables of large spatiotemporal scale (spatial scale as continental or subcontinental; temporal scale as seasonal or monthly). However, for a smaller spatiotemporal scale, the influence of local meteorological variables might be equally important. In this study, basin-scale short-term (weekly) streamflow is investigated for a possible influence of the large-scale circulation patterns and local meteorological variables on it. The ENSO and EQUINOO are used as the large-scale circulation patterns, which are established to be important for Indian hydroclimatology. On the other hand, OLR is used as local meteorological information along with streamflow information from previous time steps. The GP, which is a genetic algorithm–based approach, is used to capture the relationship between inputs and output. Different combinations of historical and previous time step streamflow, local meteorological input (OLR), and large-scale circulation pattern (ENSO and EQUINOO) were explored for the weekly basin-scale streamflow prediction. The prediction performances using 1) only the historical record, which may mimic the traditional approach; 2) only the large-scale circulation information; and 3) only the local meteorological information through OLR, are carried out. However, improvement in streamflow predictions was accomplished by the simultaneous use of the streamflow at previous time steps, large-scale circulation information (ENSO and EQUINOO), and outgoing longwave radiation along with the historical streamflow values.

It is also shown in this study that OLR information over the river basin can be successfully used to obtain better forecasts of the basin-scale streamflow for medium to large river basins. Being the proxy of rainfall, OLR serves as one of the important and reasonably influential inputs, for modeling the streamflow process. This study, thus, indicates that the effect of both large-scale atmospheric circulation patterns as well as local meteorological and hydrological influences is necessary to capture the variations in the basin-scale streamflow.

Finally, the efficacy of the GP approach to capture the complex relationship can also be mentioned. Generally, with an increase in the number of inputs, computational complexity is increased drastically for traditional models. However, as demonstrated in this study, the GP can extract the inherent information from the input set and provide the prediction with reasonable accuracy, even for such a complex system and with weekly temporal resolution.

Similar models can be developed for the other river basins by using the same approach to capture the streamflow variation. However, the number of lags for different input variables may change from basin to basin, depending upon the geographical position and the extent of the river basin. Also, the streamflow records should be free from manmade causes, such as, reservoir operation at major and medium dams, if present on the upstream side of the stream gauging site. The observed streamflow values in such cases will have to be modified in such cases, before developing the model.

Acknowledgments. This work is supported by the SERC Division, Department of Science and Technology, Government of India, through Project SR/FTP/ETA-26/07.

REFERENCES

Arkin, P. A., A. Krishna Rao, and R. Kelkar, 1989: Large-scale precipitation and outgoing longwave radiation from INSAT-1B during the 1986 Southwest monsoon season. *J. Climate,* **2,** 619–628.

Ashok, K., Z. Guan, and T. Yamagata, 2001: Impact of Indian Ocean dipole on the relationship between the Indian monsoon rainfall and ENSO. *Geophys. Res. Lett.,* **28,** 4499–4502.

——, ——, N. Saji, and T. Yamagata, 2004: Individual and combined effect of ENSO and Indian ocean dipole on the Indian summer monsoon. *J. Climate,* **17,** 3141–3155.

Barton, S. B., and J. A. Ramirez, 2004: Effects of El Niño–Southern Oscillation and Pacific Interdecadal Oscillation on water supply in the Columbia River Basin. *J. Water Resour. Plan. Manage.,* **130,** 281–289.

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

Chau, K. W., 2002: Calibration of flow and water quality modeling using genetic algorithms. *Lect. Notes Comput. Sci.,* **2557,** 720.

Cheng, C. T., C. P. Ou, and K. W. Chau, 2002: Combining a fuzzy optimal model with a genetic algorithm to solve multiobjective rainfall-runoff model calibration. *J. Hydrol.,* **268,** 72–86.

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

——, S. L. Zhou, and T. A. McMahon, 2003: Use of seasonal streamflow forecasts in water resources management. *J. Hydrol.,* **270,** 135–144.

Chowdhury, M. R., and N. Ward, 2004: Hydro-metrological variability in the Greater Ganges-Brahmaputra-Meghna Basins. *Int. J. Climatol.,* **24,** 1495–1508.

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

Dawson, C. W., and R. Wilby, 1998: An artificial neural network approach to rainfall-runoff modeling. *Hydrol. Sci. J.,* **43,** 47–66.

Douglas, W. W., S. A. Wasimi, and S. Islam, 2001: The El Niño Southern Oscillation and long-range forecasting of flows in ganges. *Int. J. Climatol.,* **21,** 77–87.

Dracup, J. A., and E. Kahya, 1994: The relationship between U. S. streamflow and La Niña events. *Water Resour. Res.,* **30,** 2133–2141.

Eltahir, E. A. B., 1996: El Niño and the natural variability in the flow of the Nile River. *Water Resour. Res.,* **32,** 131–137.

Frankone, F. D., 1998: Discipulus owner's manual, fast genetic programming based on AIML technology. RML Rep., 196 pp. [Available online at http://www.rmltech.com/.]

Gadgil, S., P. N. Vinayachandran, and P. A. Francis, 2003: Droughts of tshe Indian Summer Monsoon: Role of clouds over the Indian Ocean. *Curr. Sci.,* **85,** 1713–1719.

——, ——, ——, and S. Gadgil, 2004: Extremes of the Indian Summer Monsoon rainfall, ENSO, and equatorial Indian Ocean Oscillation. *Geophys. Res. Lett.,* **31,** L12213, doi:10.1029/2004GLO19733.

Gairola, R. M., and T. N. Krishnamurti, 1992: Rain rates based on SSM/I, OLR, and raingauge data sets. *Meteor. Atmos. Phys.,* **50,** 165–174.

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

Hong, Y.-S., and M. R. Rosen, 2002: Identification of an urban fractured rock aquifer dynamics using an evolutionary self-organizing modeling. *J. Hydrol.,* **259,** 89–104.

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

Hsu, K. L., H. V. Gupta, and S. Sorooshian, 1995: Artificial neural network modeling of the rainfall-runoff process. *Water Resour. Res.,* **31,** 2517–2530.

Jain, S., and U. Lall, 2001: Floods in a changing climate: Does the past represent the future? *Water Resour. Res.,* **37,** 3193–3205.

Jayawardena, A. W., N. Muttil, and T. M. K. G. Fernando, 2005: Rainfall-runoff modelling using genetic programming. *Proceedings of International Congress on Modelling and Simulation,* A. Zerger and R. M. Argent, Eds., Modelling and Simulation Society of Australia and New Zealand (MODSIM 2005), 1841–1847.

Kane, R. P., 1998: Extremes of the ENSO phenomenon and Indian summer monsoon rainfall. *Int. J. Climatol.,* **18,** 775–791.

Koza, J. R., 1992: *Genetic Programming: On the Programming of Computers by Means of Natural Selection.* MIT Press, 840 pp.

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

Li, T., Y. S. Zhang, C. P. Chang, and B. Wang, 2001: On the relationship between Indian Ocean sea surface temperature and Asian summer monsoon. *Geophys. Res. Lett.,* **28,** 2843–2846.

Liebmann, B., J. A. Marengo, J. D. Glick, V. E. Kousky, I. C. Wainer, and O. Massambani, 1998: A comparison of rainfall, outgoing longwave radiation, and divergence over the Amazon basin. *J. Climate,* **11,** 2898–2909.

Liong, Y. S., W. H. Lim, and G. N. Paudyal, 2000: River stage forecasting in Bangladesh: Neural networks approach. *J. Comput. Civ. Eng.,* **14,** 1–8.

Maity, R., and D. Nagesh Kumar, 2006a: Bayesian dynamic modeling for monthly Indian summer monsoon rainfall using ENSO and EQUINOO. *J. Geophys. Res.,* **111,** D07104, doi:10.1029/2005JD006539.

——, and ——, 2006b: Hydroclimatic association of monthly summer monsoon rainfall over India with large-scale atmospheric circulation from tropical Pacific Ocean and Indian Ocean region. *Atmos. Sci. Lett. Roy. Meteor. Soc.,* **7,** 101–107, doi:10.1002/asl.141.

——, and ——, 2008a: Basin-scale streamflow forecasting using the information of large-scale atmospheric circulation phenomena. *Hydrol. Processes,* **22,** 643–650, doi:10.1002/hyp.6630.

——, and ——, 2008b: Probabilistic prediction of hydroclimatic variables with nonparametric quantification of uncertainty. *J. Geophys. Res.,* **113,** D14105, doi:10.1029/2008JD009856.

——, ——, and R. S. Nanjundiah, 2007: Review of hydroclimatic teleconnection between hydrologic variables and large-scale atmospheric circulation patterns with Indian perspective. *ISH J. Hydraul. Eng.,* **13,** 77–92.

Makkeasorn, A., N. B. Chang, and X. Zhou, 2008: Short-term streamflow forecasting with global climate change implications—A comparative study between genetic programming and neural network models. *J. Hydrol.,* **352,** 336–354.

Marcella, M. P., and E. A. B. Eltahir, 2008: The hydroclimatology of Kuwait: Explaining variability of rainfall at seasonal and interannual timescales. *J. Hydrometeor.,* **9,** 1095–1105.

Minns, A. W., and M. J. Hall, 1996: Artificial neural networks as rainfall-runoff models. *Hydrol. Sci. J.,* **41,** 399–418.

Nageswara Rao, G., 1998: Interannual variation of monsoon rainfall in Godavari River Basin—Connections with the Southern Oscillation. *J. Climate,* **11,** 768–771.

Olivera, R., and D. P. Loucks, 1997: Operating rules for multireservoir systems. *Water Resour. Res.,* **33,** 839–852.

Ozelkan, E. C., and L. Duckstein, 2001: Fuzzy conceptual rainfall-runoff models. *J. Hydrol.,* **253,** 41–68.

Parthasarathy, B., H. F. Diaz, and J. K. Eischeid, 1988: Prediction of All India summer monsoon rainfall with regional and large-scale parameters. *J. Geophys. Res.,* **93,** 5341–5350.

Patri, S., 1993: Data on flood control operation of Hirakud dam. Report, Department of Irrigation, Government of Orissa, India, 270 pp. [Available from Executive Engineer, Main Dam Division, Dept. of Irrigation, Gov. of Orissa, P.O. Burla, Sambalpur, Orissa, India.]

Piechota, T. C., J. A. Dracup, and R. G. Fovell, 1997: Western U.S. streamflow and atmospheric circulation patterns during El Niño–Southern Oscillation. *J. Hydrol.,* **201,** 249–271.

Raman, H., and N. Sunilkumar, 1995: Multivariate modeling of water resources time series using artificial neural networks. *Hydrol. Sci. J.,* **40,** 145–163.

Rasmusson, E. M., and T. H. Carpenter, 1983: The relationship between eastern equatorial Pacific sea surface temperature and rainfall over India and Sri Lanka. *Mon. Wea. Rev.,* **111,** 517–528.

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

Savic, D. A., G. A. Walters, and J. W. Davidson, 1999: A genetic programming approach to rainfall-runoff modeling. *Water Resour. Manage.,* **13,** 219–231.

Wang, P. H., P. Minnis, B. A. Wielicki, T. Wong, and L. B. Vann, 2002: Satellite observations of long-term changes in tropical cloud and outgoing longwave radiation from 1985 to 1998. *Geophys. Res. Lett.,* **29,** 1397, doi:10.1029/2001GL014264.

Wang, Q. J., 1991: The genetic algorithm and its application to calibrating conceptual rainfall-runoff models. *Water Resour. Res.,* **27,** 2467–2471.

Wardlaw, R., and M. Sharif, 1999: Evaluation of genetic algorithms for optimal reservoir system operation. *J. Water Resour. Plan. Manage.,* **125,** 25–33.

Webster, P. J., and C. Hoyos, 2004: Prediction of monsoon rainfall and river discharge on 15–30-day timescale. *Bull. Amer. Meteor. Soc.,* **85,** 1745–1765.

——, A. M. Moore, J. P. Loschnigg, and R. R. Leben, 1999: Coupled oceanic–atmospheric dynamics in the Indian Ocean during 1997–98. *Nature,* **401,** 356–360.

Xie, P., and P. A. Arkin, 1998: Global monthly precipitation estimates from satellite-observed outgoing longwave radiation. *J. Climate,* **11,** 137–164.

Xiong, L., A. Y. Shamseldin, and K. M. O'Connor, 2001: A non-linear combination of the forecast of rainfall-runoff models by the first-order Takagi-Sugeno fuzzy system. *J. Hydrol.,* **254,** 196–217.

Yu, P. S., C. J. Chen, and S. J. Chen, 2000: Application of gray and fuzzy methods for rainfall forecasting. *J. Hydrol. Eng.,* **5,** 339–345.