Importance analysis of local and global climate inputs for basin-scale streamflow prediction

Rajib Maity[1] and S. S. Kashid[2]

Received 9 July 2010; revised 1 September 2011; accepted 27 September 2011; published 4 November 2011.

[1] Basin-scale streamflow is influenced by numerous local and global climate inputs. In this paper, genetic programming (GP) is combined with "importance analysis" to identify the important global climate inputs and local meteorological variables needed for prediction of weekly streamflow at the basin scale. The analysis is carried out for the Mahanadi River in India using global climate inputs, namely, the El Niño–Southern Oscillation (ENSO) index and equatorial Indian Ocean Oscillation (EQUINOO) index; local meteorological inputs, including outgoing longwave radiation (OLR), total precipitable water (TPW), temperature anomaly (TA), and pressure anomaly (PA); and streamflow information from previous time steps. The rainfall information over the basin is intentionally not utilized so that the procedure may be applicable to basins with little or no rain gauge information and to achieve a longer prediction lead time. The Birnbaum importance measure is used to assess the importance of each input. Results of this study show that the relative importance of individual input variables is influenced by time lags. It is observed that among various local meteorological inputs, OLR and PA are more important than TA and TPW. Among large-scale circulation indices, ENSO index is important for previous 5th to 7th week, whereas EQUINOO index is important for previous 3rd to 6th week. On the basis of their importance measures, 15 indices were selected from the initial group of 30 indices. The GP-derived streamflow forecasting models could predict weekly streamflow with good accuracy (correlation coefficient $r = 0.821$) for such a complex system.

Citation: Maity, R., and S. S. Kashid (2011), Importance analysis of local and global climate inputs for basin-scale streamflow prediction, *Water Resour. Res., 47*, W11504, doi:10.1029/2010WR009742.

1. Introduction

[2] Streamflow is an important and easily accessible source of water in river basins. Seasonal streamflow forecasts are useful for deciding the storage and release schedules of multipurpose and flood control reservoirs. Accurate medium-range (a few weeks in advance) streamflow forecasts are needed for the operation of reservoir systems in a river basin to help water managers make judicious water allocations for various purposes, such as irrigation, hydropower, industry, and domestic use, and for maintenance of minimum streamflows. However, the variability in streamflow discharges poses challenges to basin-scale water management.

[3] Different types of models have been developed to describe the rainfall-runoff processes and for runoff estimation over a watershed. Since intraseasonal streamflow variability is affected by both local and global hydrometeorological drivers [*Eltahir*, 1996; *Piechota et al.*, 1997; *Chiew et al.*, 1998; *Douglas et al.*, 2001; *Chandimala and Zubair*, 2007; *Maity and Nagesh Kumar*, 2008a; *Maity and Kashid*, 2010; *Kashid et al.*, 2010], streamflow prediction should rely on relevant variables representing the effects of these drivers. With the advent of various meteorological satellites, it is now possible to track the status of such local and global drivers of streamflow in a river basin.

[4] Numerous input variables influence streamflow to varying degrees. Moreover, the role of inputs varies in both space and time. Traditionally, the selection of predictors has been accomplished by regression or cross-correlation analysis. However, such methods depend on some parametric assumption of data sets and the predefined nature of correspondence (e.g., linear association for correlation coefficient). Scale-free measures, such as Kendall's tau and Spearman's rho, are free from any specific parametric assumption. However, these methods generally evaluate only pairwise association between variables. Even if these measures pass statistical significance tests, they may not reveal the true nature of dependence between input and output variables if the individual relationships (between the target variable and any input variable) are coupled with each other. In hydroclimatological analysis, the pairwise relationship between the target hydrologic variable and various hydroclimatic inputs is invariably interdependent. While some variables may show prominent association, e.g., rainfall and runoff, other factors may influence runoff generation, even though the association between them may not be statistically significant. In such cases, influencing factors are initially hypothesized, and possible interrelationships are modeled subsequently.

[1]Department of Civil Engineering, Indian Institute of Technology Kharagpur, Kharagpur, India.
[2]Department of Civil Engineering, Walchand Institute of Technology, Solapur, India.

0043-1397/11/2010WR009742

[5] To sift through a set of candidate input variables (local and global), we first employ genetic programming (GP) to check the performance of different input variable combinations, followed by "importance analysis" to evaluate the relative importance of each input variable using the Birnbaum Importance Measure (BIM). Being a data-driven machine learning evolutionary algorithm, GP does not require a priori prescription of a functional relationship. Further, the importance analysis approach considers all the potential input variables simultaneously to evaluate their relative importance. Thus, the combined influence of a group of inputs is implicitly accounted for in this method.

[6] The usefulness of climatic inputs for analysis and forecasting of hydrologic time series has been established in recent years. Specifically, the natural variation of hydrologic variables has been linked to 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; *Maity et al.*, 2007; *Marcella and Eltahir*, 2008; *Maity and Nagesh Kumar*, 2008b].

1.1. Influence of Global Climate Parameters

[7] The influence of the large-scale atmospheric circulation patterns El Niño–Southern Oscillation (ENSO) and Equatorial Indian Ocean Oscillation (EQUINOO) is discussed briefly in sections 1.1.1 and 1.1.2.

1.1.1. El Niño–Southern Oscillation

[8] El Niño–Southern Oscillation is a coupled oceanic-atmospheric mode of the tropical Pacific Ocean. The effect of ENSO on streamflow variation has been well established for different parts of the world [*Dracup and Kahya*, 1994; *Eltahir*, 1996; *Jain and Lall*, 2001] as well as over the Indian subcontinent [*Douglas et al.*, 2001; *Chowdhury and Ward*, 2004; *Chandimala and Zubair*, 2007]. *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 in the flow of the Nile River in Egypt. *Piechota et al.* [1997] explored the link between western U.S. streamflow and atmospheric circulation patterns during ENSO years. *Chiew et al.* [1998] discussed the effect of ENSO on Australian rainfall, streamflow, and droughts. The effects of El Niño–Southern Oscillation 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 on the basis of 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.*, 2003, 2004; *Maity and Nagesh Kumar*, 2006a, 2006b]. *Douglas et al.* [2001] attempted long-range forecasting of flows in the Ganges on the basis of ENSO information. *Chowdhury and Ward* [2004] studied the effect of ENSO on streamflows for the Greater Ganges-Brahmaputra-Meghna basin. *Nageswara Rao* [1997] studied interannual variation of monsoon rainfall in the Godavari River basin to establish its connection with ENSO. *Webster and Hoyos* [2004] developed a prediction scheme for monsoon rainfall and river discharge on 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.

1.1.2. Equatorial Indian Ocean Oscillation

[9] The Equatorial Indian Ocean Oscillation is the atmospheric component of the Indian Ocean dipole mode [*Saji et al.*, 1999] that is observed over the Indian Ocean [*Gadgil et al.*, 2003, 2004]. The effect of EQUINOO on rainfall and streamflow was established recently and was found to be significant for the Indian subcontinent [*Ashok et al.*, 2004; *Gadgil et al.*, 2004; *Maity and Nagesh Kumar*, 2006a, 2006b, 2008a]. The effect of EQUINOO is measured in terms of the equatorial zonal wind index (EQWIN), which is defined as the 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).

[10] The southwest monsoon is responsible for nearly 80% of Indian summer monsoon rainfall. The interaction between climates over various oceans and continents regulates the amount and distribution of the rainfall and streamflows over the subcontinent. Researchers emphasize that the strength of hydroclimatic teleconnection decreases for smaller spatiotemporal scales [*Kane*, 1998]. However, significant influence still exists at the subdivisional scale at most of the geographical locations, but the nature of the relationship varies for different subdivisions and different seasons [*Kane*, 1998; *Maity and Nagesh Kumar*, 2006b]. Therefore, it is expected that basin-scale hydrologic variables over large river basins are influenced by large-scale atmospheric circulation patterns along with local meteorological variables.

1.2. Influence of Local Meteorological Variables

[11] Basin-scale averages of local meteorological variables, namely, outgoing longwave radiation (OLR), total precipitable water (TPW), temperature, and pressure, are likely to influence the rainfall and runoff over the river basin. The effects of these variables on basin-scale streamflow are discussed in sections 1.2.1 through 1.2.3.

1.2.1. Outgoing Longwave Radiation

[12] Outgoing longwave radiation is the energy leaving the Earth in the form of infrared radiation at low energy levels. Deep clouds in largely cumulus-convection-dominated regions correspond to more intense precipitation. Summer rainfall in the tropics is usually associated with organized convective clouds, and these clouds modulate the OLR observed from satellite sensors. Hence, OLR is used as a proxy for rainfall data over the catchment for basin-scale streamflow prediction.

[13] *Gairola and Krishnamurti* [1992] have outlined a procedure for obtaining rainfall rates from a mix of satellite and surface-based observations. *Liebmann et al.* [1998] established that observed rainfall is best correlated with OLR at 10–30 day scales. The anomaly of OLR exhibits a negative correlation with precipitation over most of the globe [*Xie and Arkin*, 1998]. Rainfall estimates from OLR anomalies were reasonably consistent in both magnitude and distribution over India, with climatological mean fields derived from rain gauge measurements [*Arkin et al.*, 1989]. *Haque and Lal* [1991] discussed the space and time

variability analyses of the Indian monsoon rainfall as inferred from satellite-derived OLR data.

[14] This study deals with the Mahanadi basin in the Indian subcontinent, where high positive OLR anomaly is observed during the winter and premonsoon months and a moderate negative anomaly is observed 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. Successful forecasts of the monsoon rainfall would therefore necessitate a detailed knowledge of OLR with its spatial and temporal variability over the region.

1.2.2. Total Precipitable Water

[15] Another important parameter related to the water vapor in the atmosphere is TPW, which is currently being obtained from satellite and radiosonde measurements. TPW is the total atmospheric water vapor contained in a vertical column of unit cross-sectional area extending from the surface of the Earth to the top of the atmosphere. TPW is commonly expressed in terms of the height to which water would stand if completely condensed and collected in a vessel of the same unit cross section.

[16] Many researchers have used precipitable water information for estimating precipitation. *Xiao et al.* [2000] incorporated the Special Sensor Microwave Imager (SSM/I)-derived TPW and rainfall rate into a numerical model for studying the evolution and structure of a rapidly intensifying marine cyclone observed during Intensive Observing Period 4 (IOP 4; 4–5 January 1989) of the Experiment on Rapidly Intensifying Cyclones over the Atlantic (ERICA) and the North Atlantic Ocean. They stated that assimilation of precipitable water and rainfall rate significantly improved cyclone prediction, reflected mostly in the cyclone's track, the associated frontal structure, and precipitation along the front.

[17] SSM/I data products are produced as part of NASA's Pathfinder program. Remote Sensing Systems generate SSM/I data products using a unified, physically based algorithm to simultaneously retrieve ocean wind speed (at 10 m), water vapor, cloud water, and rain rate. *Hou et al.* [2000] assimilated SSM/I-derived surface rainfall and TPW for improving the Goddard Earth Observing System (GEOS) analysis for climate studies. They stated that assimilating rainfall data improves cloud distributions and the cloudy-sky radiation, while assimilating TPW data reduces moisture bias in the lower troposphere to improve the clear-sky radiation. They also stated that assimilation of satellite-derived precipitation and TPW can reduce state-dependent systematic errors in the OLR, clouds, surface radiation, and the large-scale circulation in the assimilated data set.

[18] *Uvo et al.* [2001] used singular value decomposition to investigate the potential of the (original and derived) point values of variables to serve as predictors for the Chikugo River basin rainfall. It was found that zonal and meridional wind speeds at 850 hPa and precipitable water are the most correlated to the Chikugo River basin rainfall and thus the most efficient predictors.

[19] *Falvey and Beavan* [2002] studied the potential of GPS precipitable water to improve mesoscale model retrievals of orographic precipitation for a prolonged rainfall event observed during the 1996 Southern Alps Experiment (SALPEX'96). They found that the assimilation of hourly GPS PW data resulted in a statistically significant (at the 3% level) improvement in the retrieved total rainfall on the upwind side of the mountain range when compared with rainfall from a control simulation that did not involve data assimilation. The best results were obtained when GPS PW data were used along with radiosonde temperature profiles. They clearly demonstrated the sensitivity of orographic rainfall to upwind precipitable water. However, they stated that the impact of precipitable water measurements in a forecast model will depend upon the length of time the water vapor information remains within the region of interest. *Olsson et al.* [2004] used artificial neural networks for rainfall forecasting by atmospheric downscaling. The experiments focused on estimating 12 h mean rainfall in the Chikugo River basin, Kyushu Island, southern Japan, from large-scale values of wind speeds at 850 hPa and precipitable water. The results indicated that longer data series are required to reproduce variability and intensity categorization that may be useful for probabilistic forecasting. They also showed that the overall performance in this region is better during winter and spring than during summer and autumn.

[20] *Ramirez et al.* [2005] used an artificial neural network technique for rainfall forecasting applied to the Sao Paulo region. They used meteorological variables from the Eta model, i.e., potential temperature, vertical component of the wind, specific humidity, air temperature, precipitable water, relative vorticity, and moisture divergence flux, as input data, keeping generated rainfall forecast for the next time step as the output.

[21] *Nezlin and Stein* [2005] studied spatial and temporal patterns of remotely sensed and field-measured rainfall in southern California. They evaluated the daily 1° resolution remotely sensed atmospheric precipitation data provided by global precipitation climatology. They analyzed data from the watersheds of southern California during the period 1996–2003, focusing on the comparison of patterns of spatial, seasonal, and interannual rainfall dynamics. They stated that precipitable water concentration measured by satellites is not always highly correlated to rainfall reaching the Earth's surface.

[22] The studies above advocate the use of TPW along with some local meteorological information for rainfall estimation. Hence, TPW is used along with OLR and other local meteorological inputs for weekly streamflow prediction in this study.

1.2.3. Temperature, Pressure, and Streamflow Information From Previous Time Steps

[23] Temperature and pressure are also important meteorological information for basin-scale hydrology. The Indian summer monsoon is a dominant component of tropical climate variability. The monsoon circulation is primarily driven by differential land-ocean heating and subsequently the release of latent heat by condensation. The seasonal variation of solar insolation produces heating of the Asian and African continents during the Northern Hemisphere summer and cooling during the winter. Differential sensible heating results in the seasonal formation of low atmospheric pressure over the Asian and African continents and high atmospheric pressure over the surrounding oceans. The release of latent heat over southern Asia and northern Africa due to monsoon precipitation results in intensification of the low atmospheric pressure and enhances the monsoon circulation. Hence, temperature and pressure over

large river basins, such as the one for Mahanadi, are expected to play a role in rainfall distribution and streamflow generation. Streamflows at previous time steps indicate the wetness of the basin and can also be used as inputs to the model. Significant serial correlation exists in successive streamflow values.

[24] Local meteorological variables (OLR, TPW, temperature and pressure anomaly) used in this study were obtained from the Physical Sciences Division of NOAA. Details are mentioned later in section 4. These data sources also provide gridded rainfall data, but these data were not used in this study for two reasons. First, the gridded climatic data (OLR and TPW in this case) are more accurate than gridded rainfall data. Rainfall would have been a potential input if it were available with reliable accuracy. Second, the effect of rainfall on streamflow is more immediate (on the temporal scale) compared to OLR and TPW. Excluding rainfall data allows for a longer lead time in the prediction of streamflow, and the method is applicable to watersheds with little or no rain gauge information.

2. Methodology

2.1. Concept of Importance Analysis

[25] The concept of importance analysis methodology originates in reliability engineering, where a system is configured as a collection of components arranged according to a specific design to achieve desired functionality with acceptable performance. Proper functioning of the system requires a proper understanding of relative "importance" of every individual component of the system.

[26] A system configuration is termed a "series system" when all components are connected in series. Failure of one component in a series system causes failure of the entire system. On the other hand, a parallel system is one where all components are connected in parallel, and failure of one component does not cause complete failure of the full system. More complex configurations, such as "series-parallel systems" or "parallel-series systems," are possible [*Elsayed*, 1996].

[27] Let us consider a system of n components. Most of the methods of assessing the importance of individual components of the system are based on observing the performance of the system when the component is functioning properly and when it is not. Let $\mathbf{X} = (X_1, X_2, X_3, \ldots, X_n)$, be the random vector representing the state of components, where X_i is the random variable denoting the ith component and X_i may take either of two states representing the component i working or not for $i = 1, 2, 3, \ldots, n$. Let X_i denote that ith unit is working, and let $\bar{X}_i$ denote that it is not working. Further, let $P(X_i)$ represent the contribution resulting from the inclusion of the input X_i toward optimum system performance. Then, $P(\bar{X}_i)$ $(= 1 - P(X_i))$ is the unreliability of the input X_i. Let R denotes reliability of the system and q $(= 1 - R)$ denote the unreliability of the system.

[28] Reliability of a parallel system is determined by estimating the probability that any one path is operational. It can also be determined by estimating the unreliability of the system and then subtracting it from unity as

$$R = 1 - [P(\bar{X}_1)P(\bar{X}_2)P(\bar{X}_3)\cdots P(\bar{X}_n)]. \quad (1)$$

[29] Thus,

$$q = P(\bar{X}_1)P(\bar{X}_2)P(\bar{X}_3)\cdots P(\bar{X}_n). \quad (2)$$

2.2. Importance Analysis for the Streamflow Prediction System

[30] The streamflow prediction system consists of certain inputs (input variable set), the processing of information, and then the output (streamflow) as a result of the process. Here input variables can be considered as "components" of the system and the output, with a certain accuracy, as a result. For our study, the system and the components can be in either of two states: "working" or "not working". A working state implies that the particular input is being used in the streamflow prediction system, whereas a not working state implies that the particular input is dropped from the streamflow prediction system. "Reliability" of a particular input in the streamflow prediction system may be expressed as a measure of the contribution, given the inclusion of that input, toward optimum streamflow prediction performance. "Unreliability" is that part of the contribution that remains unachieved, given the inclusion of that same input. The state of a system depends on the state of its components. If one of the input variables (components of a system) is dropped (not working) from the input variable combination, prediction is still possible with the remaining inputs (system is working), perhaps with less efficiency; that is, the prediction might not be optimal. Analogy of the streamflow prediction model (system) is made with the parallel system for importance analysis computations and can be visualized (known as reliability graph) as a parallel system as shown in Figure 1. From this reliability graph, an analogy can be developed toward the transfer of the concepts of reliability engineering to the streamflow prediction problem.

2.3. Assessment of Importance Measure for Individual Inputs

[31] Among a set of inputs, importance measures help to identify the relative importance of each input variable according to its importance toward streamflow prediction. The importance measure of an individual input variable is computed by using the Birnbaum Importance Measure. BIM is defined as the probability that the ith component is critical to the "proper" functioning of the system [*Elsayed*, 1996]. For our prediction model, proper can be considered to be the best possible performance of the model.

[32] For a steady state system, BIM, denoted as I_B^i, of a component i is expressed as [*Birnbaum*, 1969]

$$I_b^i = \frac{\partial G(q)}{\partial q_i}, \quad (3)$$

where $G(q)$ is a function of $\mathbf{q} = (q_1, q_2, \ldots, q_n)$ and is known as the unreliability function. For the streamflow prediction model, this can be treated as a measure of unexplained variance owing to the absence of a particular input. It is further assumed that prediction performance (or unexplained variance) due to the absence of one input is independent of the same quantity due to the absence of another input. Therefore, given that the full input set is $\{X_1, X_2, X_3, \ldots, X_n\}$, let the lack in prediction performance due to the

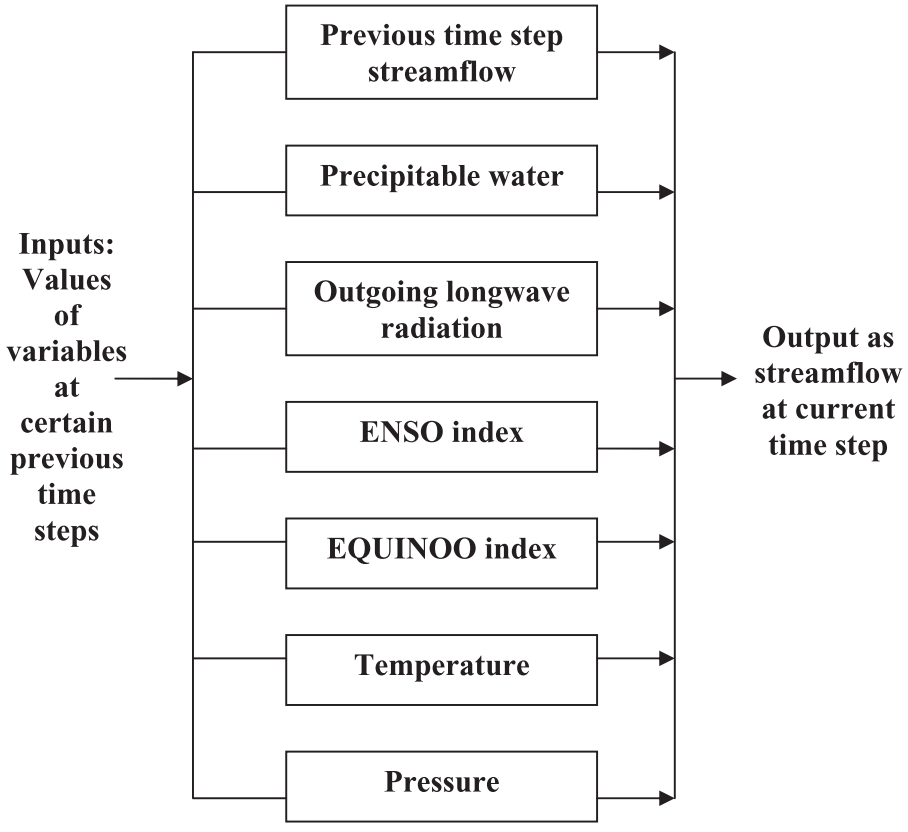

Figure 1. Reliability graph of the streamflow prediction system under study. The system can be can be treated as a parallel system as per its working style.

input set $\{X_1, X_2, \ldots, X_{r-1}, X_{r+1}, \ldots, X_n\}$ be R and that due to the input set $\{X_1, X_2, \ldots, X_{s-1}, X_{s+1}, \ldots, X_n\}$ be S; then R and S are independent of each other as long as $r \neq s$. Even though this assumption may not be valid for all possible system configurations, for the parallel system depicting the reliability graph for streamflow prediction (Figure 1), this is a valid assumption. Thus, $G(q)$ is simply expressed as

$$G(q) = q_1 q_2 \cdots q_n, \tag{4}$$

where q_i is the unreliability of the component i,

$$q_i = P(\bar{X}_i). \tag{5}$$

[33] Following equation (3), $I_b^i = q_1 q_2 \cdots q_{i-1} q_{i+1} \cdots q_n$. As q_i is defined as the probability of having a poor streamflow prediction given the ith input component, I_b^i is the probability of poor streamflow prediction without the ith component (i.e., with all other components).

2.4. Reliability Estimation From the Genetic Programming Approach: Use of Impact Frequencies

[34] Reliability (and unreliability) for a particular input is estimated from a measure of importance, known as the Birnbaum importance measure. To determine BIM, various streamflow prediction models and programs are evolved by genetic programming (GP) through machine learning. GP is basically a genetic algorithm (GA) applied to a population of computer programs. While GA usually operates on (coded) strings of numbers, GP operates on computer programs. *Koza* [1992] defined GP as a domain-independent problem-solving approach, where computer programs are evolved to approximately solve problems on the basis of the Darwinian principles of reproduction and survival of the fittest.

[35] The four general genetic operators used in GA are crossover, reproduction, mutation, and inversion. The crossover operation is a process of creating offspring programs by omitting the crossover fragment of the first parent program and then inserting the crossover fragment of the second parent program. Reproduction is the operation of just copying a program from the current population to the next population, while mutation is the process of replacing an operative function by another function and creating a new offspring program. *Koza* [1992] considered crossover and reproduction to be the two foremost genetic operations that are responsible for the genetic diversity in the population of programs. Mutation is relatively unimportant in genetic programming because the dynamic sizes and shapes of the individuals in the population already provide diversity. Thus, mutation can be considered to be a variation of the crossover operation. In the specific context of hydrology, these operations may be treated as interim processes to achieve the optimum programming structure to model the relation between the causal and target variables. Thus,

the aim of GP is to evolve a function that relates the input information to the output information, which is of the form

$$Y^m = f(X^n), \tag{6}$$

where X^n is an n-dimensional input vector and Y^m is an m-dimensional output vector. In the proposed study, the input vector consists of the historical average streamflow of the current week, ENSO and EQUINOO indices, the OLR anomaly, TPW, the temperature anomaly, the pressure anomaly, and observed streamflow values at certain number of previous time steps. The output consists of streamflow at the current time step. The flowchart of the genetic programming methodology is shown in Figure 2.

[36] Application of GP needs five major preparatory steps [*Koza*, 1992]. These five steps are (1) selection of the set of terminals, (2) selection of the set of primitive functions, (3) decision on the fitness measure, (4) decision on parameters for controlling the run, and (5) defining the method for designating results and the criterion for terminating a run. The GP tool Discipulus [*Francone*, 1998] is a software tool that develops computer programs to model the target output using the input variables during the training period. The developed program consists of sequences of instructions from an imperative programming language or machine language. The instructions are formed by using input variables, constants, and functions. These instructions are executed by processing line by line as in a series of multiple calculations through a series of processing steps [*Heywood and Zincir-Heywood*, 2002]. However, developed programs are normally highly nonlinear solutions [*Brameier and Banzhaf*, 2001, 2007]. Thus, the models are not physically based; rather, these are machine learning approaches. There are two types of computer programs provided by Discipulus as a solution, namely, a "program model" and a "team model." A program model, or evolved

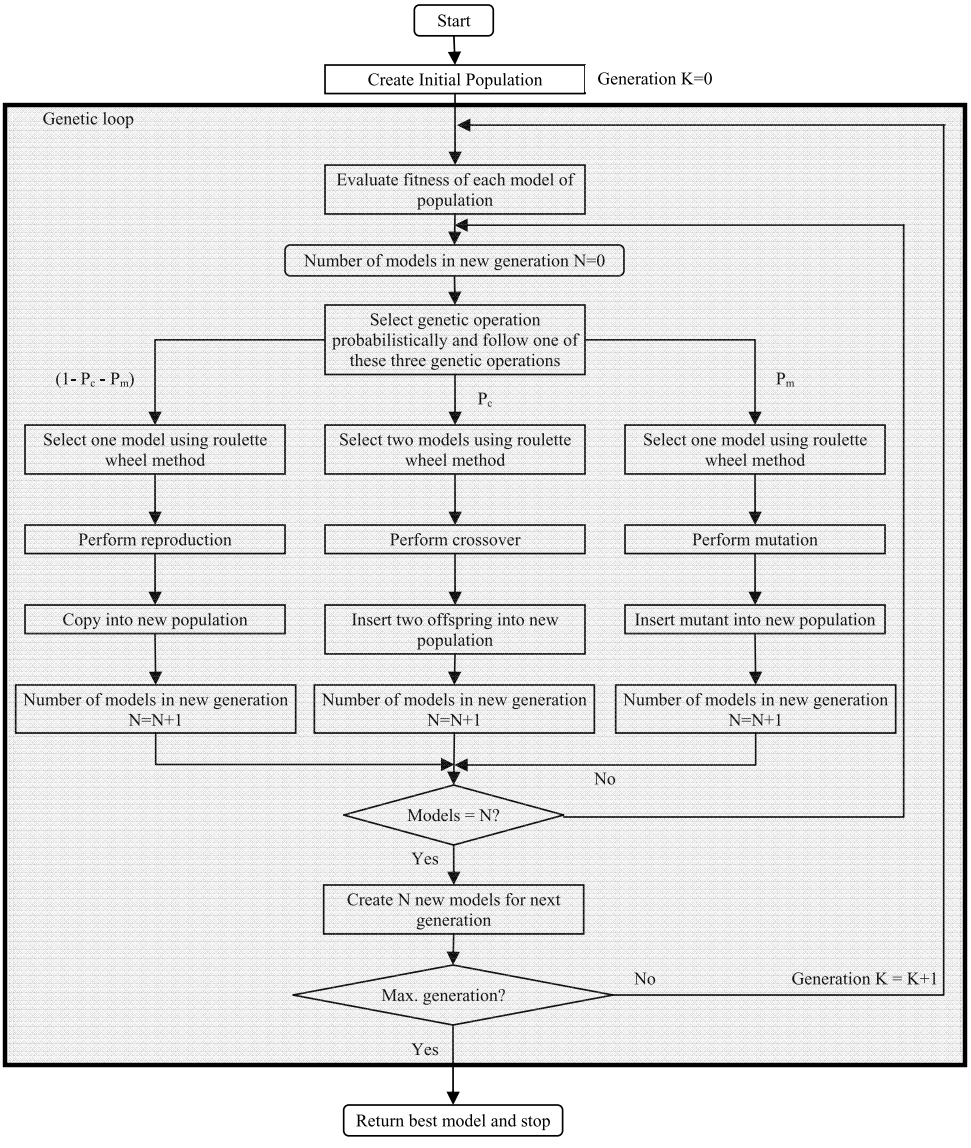

Figure 2. Flowchart for genetic programming (modified from *Hong and Bhamidimarri* [2003]), with permission from Elsevier.

program, is a single program that models the input data. A team model is a combination of a few single-program models that are combined to produce a better result than any of the single-program models. During 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. Results presented in this paper are from team models that process the data rigorously to give the best possible results.

[37] The best evolved machine code can finally be decompiled into C, Java, or a similar language. In the present study, the resulting C program was made operational for subsequent use for testing the data set in Discipulus [*Francone*, 1998].

[38] As a cautionary note, the crossover process of GP is disruptive and encourages candidate programs in the population to add unnecessary functional clauses that have little or no impact on the outcome of the function evaluation. For example, given a variable X, an extraneous clause $(X - X)$ that evaluates to 0 may be added. These extraneous clauses are adaptive in GP evolution because they lower the likelihood that crossover will disrupt the important functional clauses. Thus, it is desirable to prune the evolved candidate population at the end of a GP run to remove the extraneous clauses. Discipulus takes care of this problem by removing introns [*Francone*, 1998]. The term introns in genetic programming refers to program instructions that are included in evolved programs that have little or no effect on the output of the evolved program. The GP tool Discipulus executes a proprietary algorithm that simplifies the code by removing such instructions.

[39] After a successful run from Discipulus, it is recommended that the best 30 programs be analyzed to determine how many times each input variable appears in a way that contributes to the fitness of the programs that contain them [*Koza*, 1992; *Francone*, 1998]. The importance of every individual input variable is indicated by three indices, namely, "input impact," "average impact," and "maximum impact," yielded by the GP tool Discipulus. This tool finds how many programs out of the best 30 programs include a particular input variable. This number is treated as an indication of the contribution of the particular variable toward the prediction [*Koza*, 1992; *Francone*, 1998]. This number is then normalized to determine the input impact I_i for the ith input. Thus, a value of 1.00 for the input impact indicates that an input variable appeared in 100% of the best 30 programs. Similarly, a value of 0.67 indicates that an input variable appeared 20 times in the best 30 programs. GP has its own mechanism for computing the input impact of each variable, with important inputs being used more frequently. As a result, variables are weighted on the basis of the performance of millions of programs (models) with a large variety of variable combinations. This eliminates the need for weights to be subjectively assigned to input variables by a user.

[40] Discipulus uses the sum of squared errors for calculating the fitness of a program. It calculates the error produced by the evolved program for each example in the data set after removing all instances of that input from the group of 30 best programs of the project and replacing it with a permuted value of that input. Then, the average (and maximum) of all those errors is computed over the whole data set for that program. Next, the average (and the maximum) errors are computed for all 30 programs. It is then scaled between 0 and 1.0. The scaled value from average error provides the average impact and that from maximum error provides the maximum impact. For both, the greater the value is, the higher the impact of the removal of that particular input from the input set is. A suitable combination of the three indices (input impact, average impact, and maximum impact) may be used for computing the importance of an individual input variable.

[41] Millions of programs are evolved and tested by Discipulus; hence, there is very little difference of fitness among the best 30 programs. For example, typically, the r^2 value for the 1st program was found to be 0.55 and that for the 30th program was found to be 0.51, and as such, the best 30 programs were given the same weight. The GP tool Discipulus invariably uses the best 30 programs for calculating input impacts. The users of the tool have no choice in this. The model developer opines that the choice of the best 30 programs provides a sufficient chance for all input variables to be represented and that this group may be treated as being comprehensive. It is expected that the more sensitive variables will appear frequently in the best 30 programs.

[42] The GP will allow a particular input in these best programs depending on the overall prediction performance. So I_i may be linked with the probability of a particular input contributing to the overall performance of the programs that were evolved by GP through machine learning. $P(X_i)$ is expressed as $P(X_i) = m/(n + 1)$, where the input X_i appears in m of the total n (30) best programs and, as mentioned earlier, $I_i = m/n$. Thus, the unreliability of the ith input can be expressed as

$$q_i = P(\bar{X}_i) = 1 - P(X_i) = 1 - \frac{m}{n+1}. \tag{7}$$

[43] Once these unreliabilities are obtained, BIM can be computed as described in section 2.3. In equation (7), the normalizing factor is chosen as be $n + 1$ to accommodate the instance when a particular input appears in all the 30 best programs. To be specific, this is to ensure that $q_i \neq 0$ for all i. If it is normalized by n, then there could be a chance of getting $P(X_i)$ as 1 in the case of the inputs having input impact equal to 1. This will lead to having a nonzero BIM value only for that input and zero values for all rest inputs. On the other hand, normalizing by $n + 1$ will ensure nonzero BIM values for all the inputs without any loss of generality and importance order of the inputs. On the basis of the BIM for individual input parameters, the importance of different input parameters can be ranked in a relative fashion.

[44] Two points are worthy of noting here. First, BIM is not used to check for overparameterization as this aspect is addressed by the genetic programming used for the analysis. This is known as "parsimony pressure" in Discipulus [*Francone*, 1998]. In the evaluation process, the longer input strings (with more parameters) are automatically omitted during the evolution process. That said, the chance of overparameterization is more in the case of a small number of patterns (the combination of one target data and the corresponding input data is known as one pattern) being

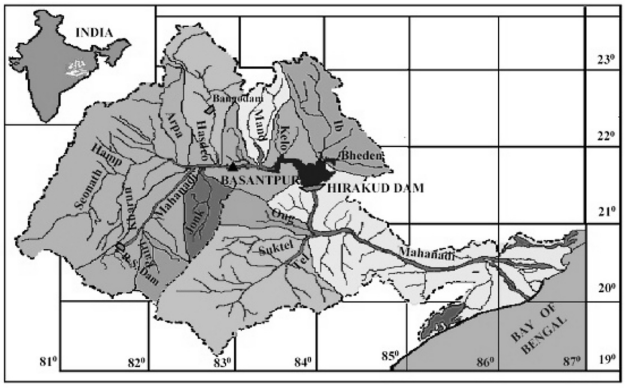

Figure 3. Location map of catchment and subbasins of Mahanadi River (source: Central Water Commission, Mahanadi Division, Burla, India).

used for calibration. Thus, reasonably large numbers of patterns are necessary for the evolution process through GP to avoid the chance of overparameterization. Second, the theoretical range of BIM varies from $G(q)$ to $+\infty$. It is a relative measure and can only determine whether the importance of an individual input is more or less when compared to another input.

[45] More details on genetic programming methodology [*Koza*, 1992; *Francone*, 1998] and its application for streamflow prediction can be found elsewhere [*Makkeasorn et al.*, 2008; *Maity and Kashid*, 2010; *Kashid et al.*, 2010].

3. Study Area

[46] The basin-scale streamflow measurements referred to in this study are observed at the Basantpur river gauging site across the Mahanadi River in India. The catchment area of the Mahanadi River at this site encompasses extensive areas of Chhattisgarh state. The Mahanadi River rises in the highlands of Chhattisgarh and flows through Orissa to discharge into the Bay of Bengal along the east coast of India. Hirakud is one of the most important multipurpose reservoirs in India, built across the Mahanadi River. The location map of the Mahanadi catchment (Figure 3) depicts the subbasins of the Mahanadi River, Hirakud Dam, and the Basantpur stream gauging station. Hirakud Dam intercepts about 83,400 km^2 of the Mahanadi catchment, which is about 65% of the total catchment area of Mahanadi. The river gauging station, Basantpur, is located just upstream of the Hirakud reservoir in the state of Orissa, India. It is operated by the Central Water Commission (CWC) of India.

4. Data

[47] Daily streamflow data at the Basantpur site were obtained from the Office of Executive Engineer, Mahanadi Division, CWC, Burla, Orissa for the period January 1990 to December 2003. The historical weekly average streamflow in the Mahanadi River at the Basantpur site is shown in Figure 4. It can be observed from Figure 4 that on average, monsoon streamflows were significant from the third week of June (approximately the 25th week in Figure 4). Following *Maity and Kashid* [2010], monsoon streamflows for 18 successive weeks, starting from the third week of June, were considered for weekly streamflow prediction.

[48] The sea surface temperature anomaly (SSTA) from the Niño 3.4 region (120°W–170°W, 5°S–5°N) was used as the ENSO index in this study. Weekly SSTA data are obtained from the Web site of the National Weather Service's Climate Prediction Center (http://www.cpc.noaa.gov/data/indices/) for the period January 1990 to December 2003. Similarly, EQWIN was used as the EQUINOO index as explained earlier in section 1.1.2. Weekly surface wind data for the period January 1990 to December 2003 were obtained from the National Centers for Environmental Prediction (http://www.cdc.noaa.gov/Datasets).

[49] The OLR data used in this study for streamflow estimation in Mahanadi basin were the daily mean values of

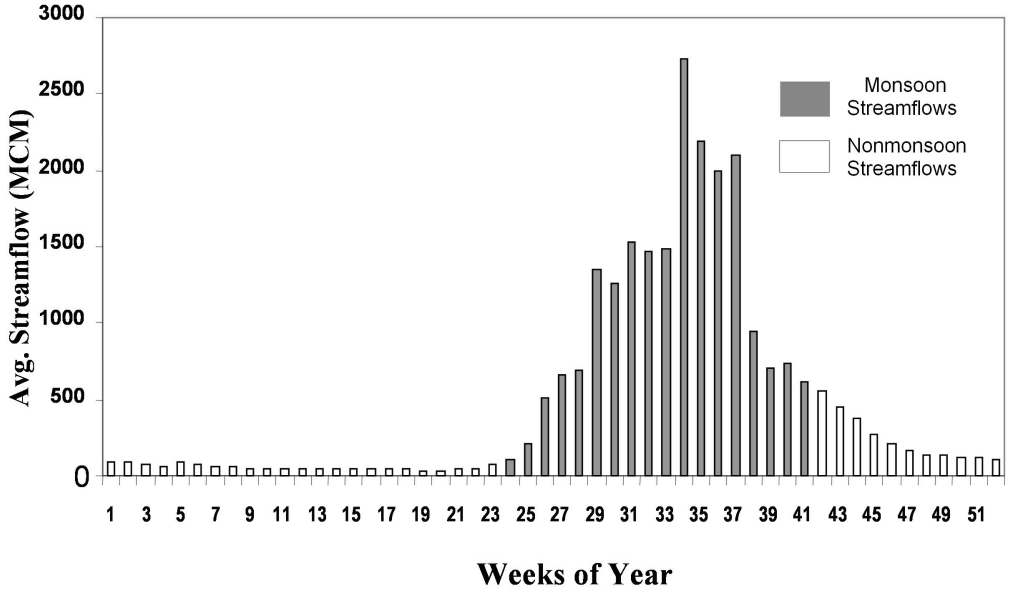

Figure 4. Historical weekly average streamflow (monsoon weeks are shaded).

OLR over the Indian region (15°N–25°N, 75°E–90°E) at 2.5° latitude and longitude intervals for 15 years from 1 January 1990 to 31 December 2004. The daily mean OLR data are used to derive weekly means. The long-term mean OLR was calculated first for a 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 average weekly OLR over the region from the observed OLR value for the particular week. The interpolated OLR data used in this study were obtained from the Climate Prediction Center Web site (http://www.cdc.noaa.gov).

[50] Similarly, daily mean values of precipitable water were obtained over the same region at 2.5° latitude and longitude intervals for the same period. The daily mean precipitable water values were used to derive weekly means. The weekly precipitable water values for the specified region were obtained by summing the weekly grid values in a region and then dividing the sum by the number of grid points composing the region.

[51] In the same way, temperature and pressure data used in this study were the daily mean values of temperature and pressure over the same region at 2.5° latitude and longitude intervals for the same period. The daily mean temperature and pressure were used to derive weekly means. The long-term means of temperature and pressure were calculated first for a particular week. The weekly mean temperature and pressure for specified region were obtained by summing the weekly grid values in the region and then dividing the sum by the number of grid points composing the region. The temperature and pressure anomalies for the region under consideration were then computed by deducting average weekly temperature and pressure over the region from the observed temperature and pressure values for the particular week.

[52] The temperature, pressure, and precipitable water data used in this study were obtained from the Climate Prediction Center Web site (http://www.cdc.noaa.gov). The catchment area of the Mahanadi River at the Basantpur stream gauging site extends between 19°N and 24°N latitudes and 80°E and 84°E longitudes. However, the local meteorological data, namely, OLR, TPW, temperature, and pressure, used in this study for streamflow estimation in the Mahanadi basin were collected from the extended region (15°N–25°N, 75°E–90°E) at 2.5° latitude and longitude intervals. The extensions beyond the catchment were deliberately included to capture the effect of cloud systems passing over a larger area across the basin and to get more grid points for fair values of weekly averages over the region. Weekly data from 1 January 1990 to 31 December 2003 were used for this study. Out of this period, weekly data from 162 monsoon weeks from years 1990 through 1998 were used for training purposes, and data from 90 monsoon weeks from years 1999 through 2003 were used for testing purpose.

5. Results and Discussions

[53] Pearson's correlation coefficients and Kendall's tau between the 30 inputs and weekly streamflow values are calculated in the first step of analysis to assess the association between the inputs and the output. Pearson's correlation coefficient is a measure of linear association, whereas Kendall's tau is a scale-free measure of association. Kendall's tau is defined as

$$\tau = \frac{n_c - n_d}{{}^{n}C_2}, \tag{8}$$

where n_c is the number of concordant pairs, n_d is the number of discordant pairs in the data set under consideration, and n is the total number of pairs. The pairs (x_i, y_i) and (x_j, y_j) are said to be concordant if $x_i > x_j$ and $y_i > y_j$ or $x_i < x_j$ and $y_i < y_j$. On the other hand, if $x_i > x_j$ and $y_i < y_j$ or $x_i < x_j$ and $y_i > y_j$, the pair is said to be discordant. The total number of pairs within a data set is ${}^{n}C_2$. A high value of τ denotes that most pairs are concordant, indicating the two rankings are consistent. The values of Kendall's tau range from -1 (100% negative association, or perfect inversion) to $+1$ (100% positive association, or perfect agreement). A value of zero indicates the absence of association. Correlation coefficients and Kendal's tau values are summarized in Table 1. As mentioned, since the input variables are dependent on each other, not all these values are statistically significant. Thus, to understand the relative impact of each input toward its contribution to explain the observed variability in streamflow, the Birnbaum importance measure is used.

5.1. Streamflow Prediction Analyses using GP

[54] Streamflow prediction analyses were performed by two different approaches. The first approach was a comprehensive analysis that included a reasonable number of relevant inputs (30 inputs) for streamflow prediction. Birnbaum importance measures were calculated for each of the 30 inputs. In the second approach, only the 15 most significant input variables were identified from the comprehensive analysis based on the Birnbaum importance measures, and streamflows were predicted with this revised set.

[55] Thus, in the first step, weekly streamflow was modeled using GP as a function of 30 input variables, namely, (1) historical average weekly streamflow for the particular week, (2) streamflow at previous weekly time steps, i.e., $t-3$ through $t-1$, (3) ENSO indices of $t-12$ through $t-2$ previous weekly time steps, (4) EQUINOO indices of $t-8$ through $t-2$ previous weekly time steps, (5) the OLR anomaly over the basin for two previous weekly time steps, (6) total precipitable water over the basin for two previous weekly time steps, (7) the temperature anomaly (TA) over the basin for two previous weekly time steps, and (8) the pressure anomaly (PA) over the basin for two previous weekly time steps (total 30 inputs).

[56] Thus, the streamflow prediction equation for basin-scale streamflow can be presented as follows:

$$\begin{aligned} \mathrm{SF}_t = f\{&\mathrm{HASF}_t, (\mathrm{SF}_{t-2}, \mathrm{SF}_{t-1}), (\mathrm{EN}_{t-12}, \ldots, \mathrm{EN}_{t-2}),\\ &(\mathrm{EQ}_{t-8}, \ldots, \mathrm{EQ}_{t-2}), (\mathrm{OLR}_{t-2}, \mathrm{OLR}_{t-1}),\\ &(\mathrm{TPW}_{t-2}, \mathrm{TPW}_{t-1}), (\mathrm{TA}_{t-2}, \mathrm{TA}_{t-1}), (\mathrm{PA}_{t-2}, \mathrm{PA}_{t-1})\}, \end{aligned} \tag{9}$$

where HASF stands for historical weekly average streamflow, SF stands for streamflow, EN stands for the ENSO index, EQ stands for the EQUINOO index, and OLR, TPW, TA, and PA are as already defined.

[57] The time lags for different inputs in the above mentioned input variable combination were based on prior knowledge of influence of these inputs on monsoon rainfall.

Table 1. Correlation Coefficients and Kendall's Tau Between Different Input Variables and Weekly Streamflow

Variable Number	Input Variable[a]	Statistical Measure	
		Correlation Coefficients	Kendall's Tau
1	HASF	0.434	0.427
2	OLR_{t-2}	−0.336	−0.361
3	OLR_{t-1}	−0.306	−0.307
4	SF_{t-3}	0.368	0.381
5	SF_{t-2}	0.389	0.457
6	SF_{t-1}	0.587	0.598
7	EN_{t-12}	0.006	−0.031
8	EN_{t-11}	0.003	−0.031
9	EN_{t-10}	0.004	0.017
10	EN_{t-9}	0.023	−0.003
11	EN_{t-8}	0.052	0.015
12	EN_{t-7}	0.071	0.033
13	EN_{t-6}	0.088	0.039
14	EN_{t-5}	0.076	0.04
15	EN_{t-4}	0.075	0.033
16	EN_{t-3}	0.062	0.015
17	EN_{t-2}	0.041	0.004
18	EQ_{t-8}	0.122	0.078
19	EQ_{t-7}	0.179	0.11
20	EQ_{t-6}	0.217	0.148
21	EQ_{t-5}	0.193	0.149
22	EQ_{t-4}	0.176	0.151
23	EQ_{t-3}	0.149	0.119
24	EQ_{t-2}	0.120	0.087
25	TPW_{t-2}	0.069	0.119
26	TPW_{t-1}	0.189	0.126
27	TA_{t-2}	−0.054	−0.091
28	TA_{t-1}	−0.134	−0.152
29	PA_{t-2}	−0.244	−0.144
30	PA_{t-1}	−0.262	−0.099

[a]HASF, historical monthly average streamflow; OLR, outgoing longwave radiation; SF, streamflow; EN, El Niño–Southern Oscillation (ENSO) index; EQ, equatorial Indian Ocean Oscillation (EQUINOO) index; TPW, total precipitable water; TA, temperature anomaly; PA, pressure anomaly.

Approximately 3 months (12 weeks) lead time was considered for ENSO and 2 months (8 weeks) for EQUINOO. The large distance of the oceans from the river basin is the basic consideration behind accounting for up to 12 week lags for ENSO indices and 8 week lags for EQUINOO indices. Such temporal lags for these indices were also useful for incorporating evolutionary trends in the data. Thus, the evolutionary trends of ENSO and EQUINOO were used to incorporate the trend of climate forcing on the hydrologic process through "hydroclimatic teleconnection."

[58] The catchment area at the Basantpur stream gauging site across the Mahanadi is spread over 83,500 km^2. This spatial extent of the river basin was the decisive factor for all local meteorological input variables, namely, OLR, TPW, TA, and PA in the comprehensive analysis. Hence, the influence of the local meteorological variables was considered for two previous weeks. The OLR and TPW lags are useful for considering future trends of precipitation. Similarly, time-lagged streamflows provide indications of streamflow trends in the weeks to come.

[59] The Pearson's correlation coefficient r between observed and predicted streamflow was computed for this "total" combination with 30 input variables. The explained variance was found to be 68.7% ($r^2 = 0.687$) during the training period and 59.9% ($r^2 = 0.599$) during the testing period. The results may be viewed graphically in Figure 5, which shows comparisons of observed and predicted streamflows during training period, and in Figure 6 for comparison between observed and predicted streamflows for the testing period. The same results are depicted in "anomaly" form in Figures 7 and 8. The input variable combinations are reported in Table 2 along with the corresponding input impacts. It is worthwhile to mention here that the effect of the large-scale inputs (e.g., ENSO) was investigated in an earlier study [*Maity and Kashid*, 2010]. It was shown that exclusion of large-scale inputs could explain 52.9% of variability. Other possible combinations (total 11) were also tried and published in earlier studies, where the gradual evolution of combinations was carried out on a trial-and-error basis [*Maity and Kashid*, 2010]. However, the methodology in this paper identifies the important inputs to achieve the optimum combination in a single step on the basis of a mathematical approach.

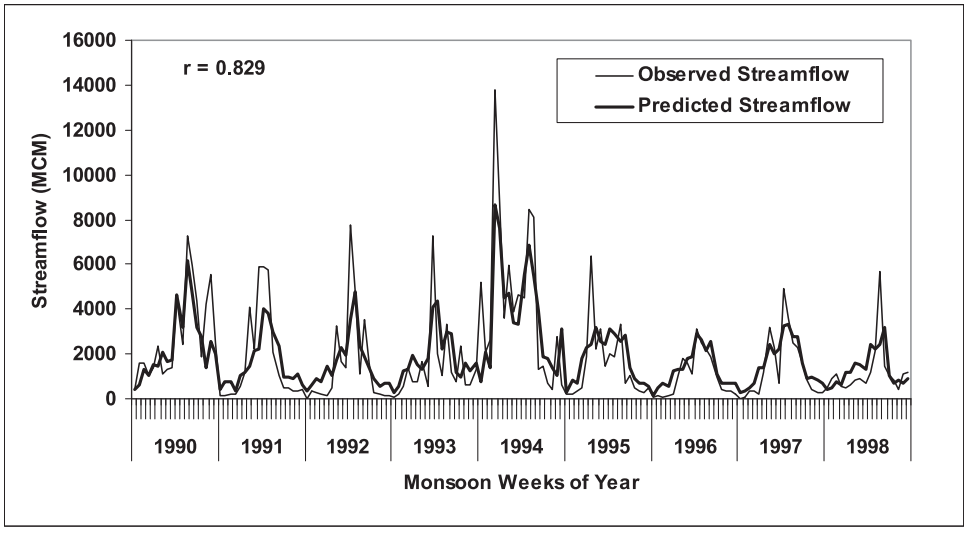

Figure 5. Comparison between observed and predicted streamflow during the training period for the comprehensive analysis (see Table 2).

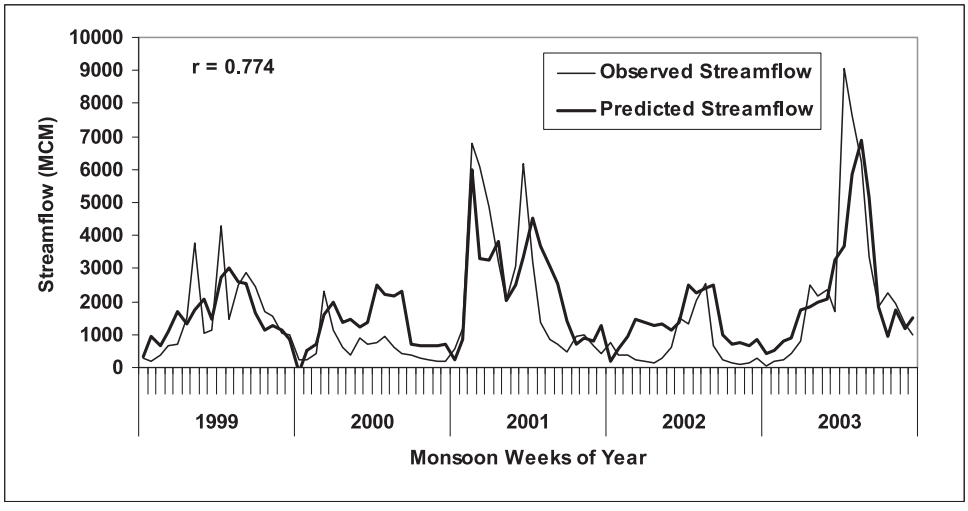

Figure 6. Comparison between observed and predicted streamflow during the testing period for the comprehensive analysis (see Table 2).

[60] It may be observed from Table 2 that the three indices, input impacts, average impacts, and maximum impacts, are concordant; that is, they rise and fall together. Hence, any one of them may be used for the BIM computation, and the input impact was chosen in this study.

[61] The genetic programming analysis is called a "project." A run of the project with the genetic programming tool Discipulus starts with a single population of evolving programs, evolves them into high-precision models, and then stops. A Discipulus project thus is a collection of many such runs, performed in series or in parallel. Generally, each of the runs in a project is performed with a different random seed and may be performed with different run parameters. Hence, even though it is possible to force Discipulus to perform a project that consists of a single run, extensive research by *Francone* [1998] has established that multiple runs are much more likely to produce good results. A "team" of programs was used to obtain the results in this study. Specifically, a total of 117 runs were performed and 678,028,397 programs were evolved by GP before arriving at the best 30 programs representing the model.

5.2. Birnbaum Importance Measures for Individual Inputs

[62] The methodology of the Birnbaum importance measures was discussed in section 2. The calculations are presented in tabular form in Table 2. The second column lists all the input variables of the comprehensive analysis for streamflow prediction. The inputs can be treated as components of the system, and they are numbered $X_1, X_2, \ldots, X_{30}$. The input impacts computed by the genetic programming tool I_i for every individual input variable are listed in the third column.

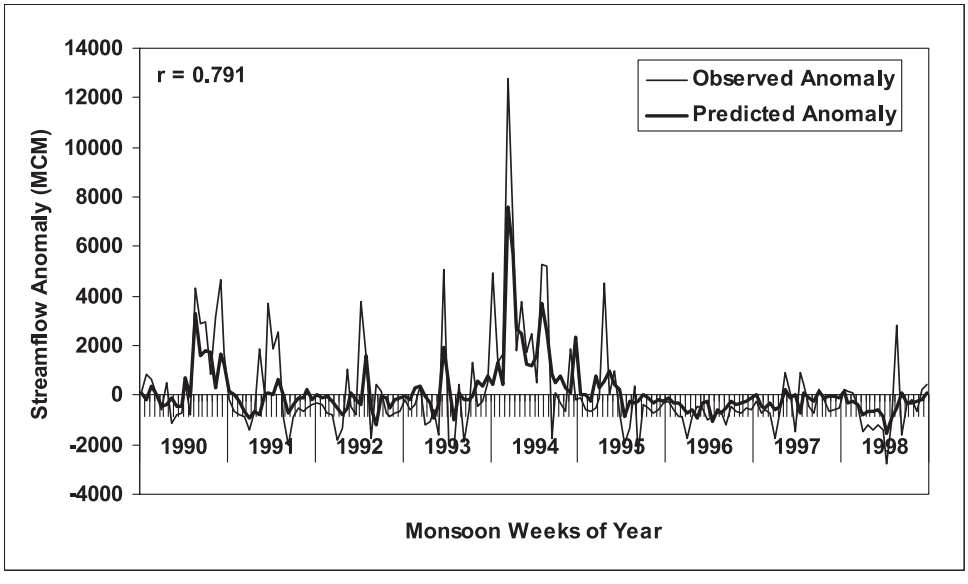

Figure 7. Comparison between the observed and predicted streamflow anomalies during the training period for the comprehensive analysis (see Table 2).

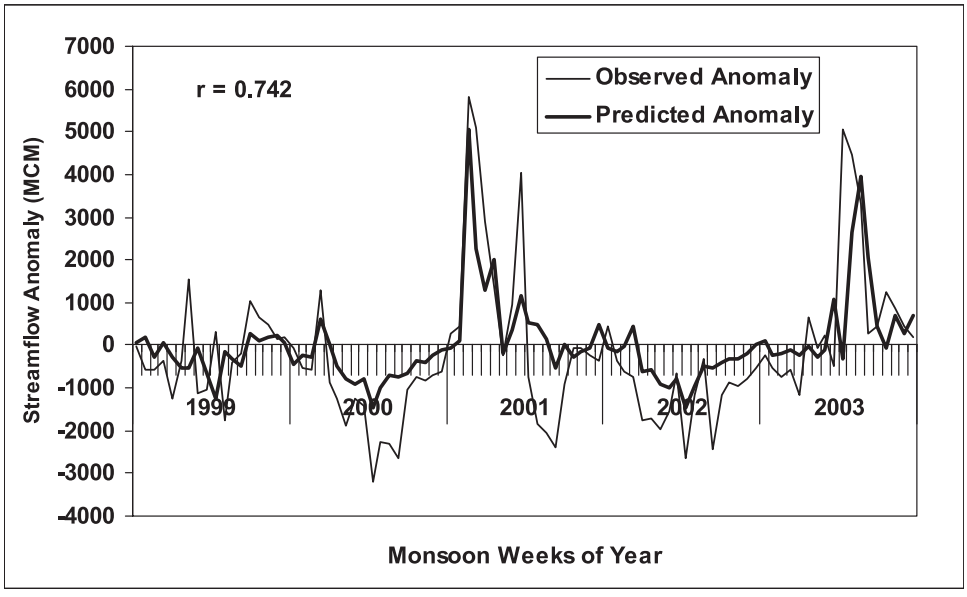

Figure 8. Comparison between observed and predicted streamflow anomalies during the testing period for the comprehensive analysis (see Table 2).

As discussed in section 2, $P(X_i)$ and q_i for each individual input were computed as per equation (7). The function $G(q)$ in equation (4) was calculated as a product of all measures of unreliability. BIM (I_b^i) of every individual component was calculated as per equation (3) and is shown in the sixth column. The BIM values reflect the relative importance of all input variables in streamflow prediction.

[63] After analyzing the Birnbaum importance measures computed for individual variables, 15 inputs were selected for a second streamflow prediction analysis based on BIM values.

Table 2. Input Impacts and Birnbaum Importance Measures for 30 Inputs[a]

Variable Number	Input Variables	Input Impacts by GP Tool I_i	Average Impact	Maximum Impact	Birnbaum Importance Measure I_b^i[b]
1	**HASF**	**0.80**	**0.42**	**0.49**	**9.02E-5**
2	**OLR_{t-2}**	**0.43**	**0.10**	**0.11**	**3.49E-5**
3	**OLR_{t-1}**	**0.33**	**0.17**	**0.36**	**2.99E-5**
4	SF_{t-3}	0.23	0.00	0.31	2.62E-5
5	**SF_{t-2}**	**0.83**	**0.20**	**0.39**	**1.04E-4**
6	**SF_{t-1}**	**0.43**	**0.33**	**0.50**	**3.49E-5**
7	EN_{t-12}	0.13	0.01	0.05	2.33E-5
8	EN_{t-11}	0.07	0.01	0.08	2.19E-5
9	EN_{t-10}	0.03	0.02	0.04	2.10E-5
10	EN_{t-9}	0.03	0.00	0.13	2.10E-5
11	EN_{t-8}	0.03	0.00	0.06	2.10E-5
12	**EN_{t-7}**	**0.37**	**0.25**	**0.39**	**3.17E-5**
13	**EN_{t-6}**	**0.20**	**0.19**	**0.25**	**2.53E-5**
14	**EN_{t-5}**	**0.37**	**0.35**	**0.46**	**3.17E-5**
15	EN_{t-4}	0.07	0.15	0.36	2.19E-5
16	EN_{t-3}	0.20	0.29	0.32	2.53E-5
17	EN_{t-2}	0.10	0.18	0.41	2.26E-5
18	EQ_{t-8}	0.13	0.14	0.33	2.33E-5
19	EQ_{t-7}	0.10	0.03	0.28	2.26E-5
20	**EQ_{t-6}**	**0.67**	**0.42**	**0.47**	**5.80E-5**
21	**EQ_{t-5}**	**0.03**	**0.17**	**0.20**	**2.10E-5**
22	**EQ_{t-4}**	**0.07**	**0.20**	**0.42**	**2.19E-5**
23	**EQ_{t-3}**	**0.27**	**0.25**	**0.43**	**2.76E-5**
24	EQ_{t-2}	0.20	0.10	0.28	2.53E-5
25	TPW_{t-3}	0.13	0.08	0.34	2.33E-5
26	**TPW_{t-1}**	**0.23**	**0.15**	**0.21**	**2.62E-5**
27	TA_{t-2}	0.20	0.13	0.19	2.53E-5
28	**TA_{t-1}**	**0.07**	**0.06**	**0.40**	**2.19E-5**
29	PA_{t-2}	0.13	0.15	0.21	2.33E-5
30	**PA_{t-1}**	**0.80**	**0.49**	**0.51**	**9.02E-5**

[a]The 15 selected inputs identified as important are shown in bold. GP, genetic programming; HASF, historical monthly average streamflow; OLR, outgoing longwave radiation; SF, streamflow; EN, ENSO index; EQ, EQUINOO index; TPW, total precipitable water; TA, temperature anomaly; PA, pressure anomaly.
[b]From equations (3) and (7).

It might be noted that the 15 selected inputs are not those with the 15 highest BIM values out of 30 inputs, taken in descending order. It was decided to include at least one value of all these inputs depending upon their importance. Thus, they are selected by giving due representation to every type of input with significant number of time steps. The selected inputs were HASF; OLR of the two previous weeks, OLR_{t-2} and OLR_{t-1}; observed streamflow of the two previous weeks, SF_{t-2} and SF_{t-1}; ENSO indices EN_{t-7}, EN_{t-6}, and EN_{t-5}; EQUINOO indices EQ_{t-6}, EQ_{t-5}, EQ_{t-4}, and EQ_{t-3}; TPW_{t-1}; TA_{t-1}; and PA_{t-1}. Thus, the basin-scale streamflow equation with the selected inputs was formulated as

$$\begin{aligned} SF_t = f\{&HASF_t, (SF_{t-2}, SF_{t-1}), (EN_{t-7}, EN_{t-6}, EN_{t-5}), \\ &(EQ_{t-6}, EQ_{t-5}, EQ_{t-4}, EQ_{t-3}), (OLR_{t-2}, OLR_{t-1}), \\ &TPW_{t-1}, TA_{t-1}, PA_{t-1}\}. \end{aligned} \quad (10)$$

[64] With this selected combination of input variables in equation (10), the explained variance was found to be 72.9% ($r^2 = 0.729$) during the training period and 67.4% ($r^2 = 0.674$) during the testing period. The improvement in the results may be attributed to the selection process adopted for the streamflow prediction methodology. Unlike parametric approaches, such as general regression and an autoregressive integrated moving average, the explained variability may not always improve (or remain same) with the increase in the number of inputs. Being a machine learning approach, this may not be always true while using genetic programming. The presence of fewer effective inputs may result in poor performance. This might be the possible reason for getting a higher r^2 value while using 15 inputs compared to that of 30 inputs. The central idea of this analysis is to find important input variables from the input set, and an improvement in performance is not unexpected even with a smaller (but more important) number of inputs.

[65] The Birnbaum importance measures provide a general platform to arrive at the conclusions. There exists a mathematical relation between input impact and BIM. However, BIM values contribute additional information in that they convey the relative importance of input parameters within the entire system. Thus, BIM provides a justification to use these measures to decide on the ordering of importance from a reliability perspective. Without this mathematical foundation it may not be logical to order the inputs according to their importance.

[66] In the way the input impact is defined in the context of GP, there exists a (direct) mathematical relation between input impact and BIM. However, to justify this, the concept of reliability engineering is used in the important analysis. In this study, the input impact is used, as obtained from a machine learning approach (GP), to represent reliability. Nonetheless, it is not always necessary that the measure of contribution is determined on the basis of the GP approach. There could be other similar (machine learning) approaches. The way to quantify the measure of the contribution toward the optimal prediction performance can be decided by the user. An example is shown in this study in the case of GP. However, to understand the relative impacts of each input toward its contributions to explain the observed variability in streamflow, the measure of BIM is used as mentioned earlier in section 2.3. It may be noted that the BIM is a relative measure of importance among different inputs. This is not an absolute measure. Thus, quantification of the significance threshold may not be possible, whereas arranging the inputs in order of their importance is possible. Thus, the results of this study show that the relative importance of individual input variable varies with different lags for different inputs.

[67] It is observed that among various local meteorological inputs, OLR and PA are more important than TA and TPW. Among large-scale circulation indices, the ENSO index is important for the previous 5th to 7th week, whereas the EQUINOO index is important for the previous 3rd to 6th week. On the basis of the importance measure, 15 inputs (8 variables with different lags) were selected from the group of initial 30 inputs (8 variables with different lags).

[68] Details of the variable combinations for the complete analysis (comprehensive analysis) and with 15 selected inputs are shown in Table 2. A comparison of observed and predicted streamflows during training and testing is given in Figures 9 and 10, respectively. The same results in anomaly form are presented in Figures 11 and 12. Residual plots and corresponding autocorrelograms are also investigated. Plots of residuals indicate the absence of any major trend, and the autocorrelograms indicate the absence of any significant serial correlation in the residuals. This is true for both training and testing periods with both the full 30 variables and the 15 selected variables as the input set. The input impacts and the Birnbaum importance measures calculated for the analysis are listed in Table 2.

[69] Inspection of the results in Tables 1 and 2 reveals a few observations. First, Kendall's tau between the pressure anomaly and streamflow is very low and does not seem to have any significance, whereas a reasonably high correlation coefficient indicates a different situation. However, BIM analysis clearly shows the pressure anomaly to be an important input. Second, in the case of the ENSO index, both the correlation coefficient and Kendall's tau are low for all the lags considered. There are a few lags (lags 4 through 8) with relatively higher values, but values are undoubtedly very insignificant. In the case of BIM analysis, ESNO indices at these lags were found to be comparable to other important inputs. Similar observations can be made for the EQUNIOO index as well. However, the correlation coefficient and Kendall's tau values are not as low as those in the case of ENSO. Last, a significantly high correlation coefficient and Kendall's tau values are obtained in the case of historical streamflow, streamflow from previous steps, and OLR. For these inputs, BIM values are also found to be high. The reason behind these observations could be the fact that both the correlation coefficient and Kendall's tau generally evaluate only pairwise association between variables. They may not reveal the true nature of dependence between input and output variables if the individual relationships are coupled with each other. Thus, these observations reveal the danger of relying only on pairwise dependence measures when the relationship between the target hydrologic variable and various inputs are expected to be interdependent.

5.3. Some Further Discussions

[70] The physical significance of the important inputs identified by this method is worth examining. Time lags for different input variables will likely change from basin to basin, depending upon the geographical position and the extent of the river basin. The catchment area at the Basantpur

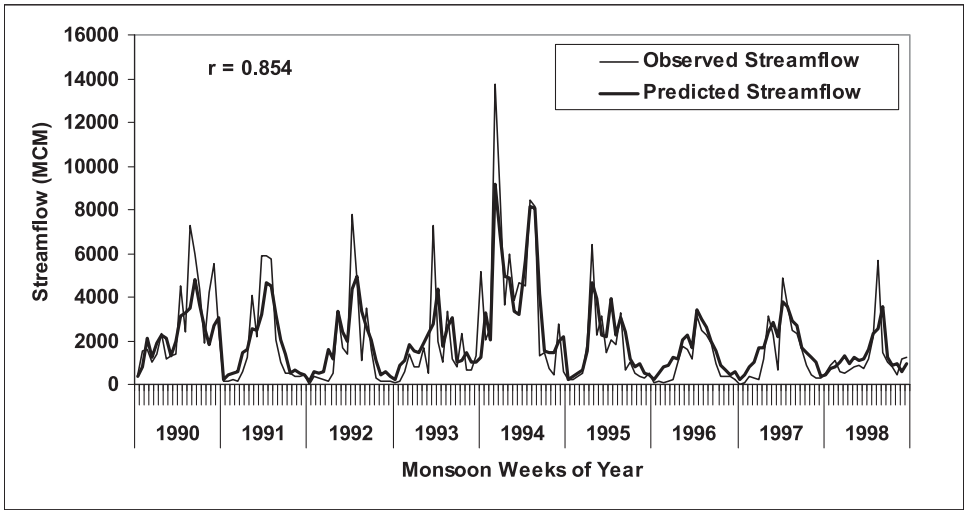

Figure 9. Comparison between observed and predicted streamflow during the training period for the analysis with 15 selected variables (see Table 2).

stream gauging site across Mahanadi is spread over 83,500 km^2. This spatial extent of the river basin is the decisive factor for all local meteorological input variables, namely, OLR, TPW, TA, and PA and the associated time lags in the comprehensive analysis. The historical streamflow provides the long-term mean status of streamflow for a week. The difference between this mean and actual observation is the deviation from the normal trend. Such deviations result from a combination of inherent stochasticity and the effect of external variables due to different hydroclimatological factors. Streamflows from previous time steps reflect the recent condition of the catchment from a hydrological point of view. OLR is intercepted by the presence of convective clouds, which may generate rainfall and streamflow; thus, OLR serves as a surrogate for rainfall and, in turn, shows a strong relationship with streamflow. Among the local inputs, OLR (meteorological) and observed streamflow with a 2 week time lag were found to be important.

[71] TPW, TA, and PA also have direct and/or indirect relation to convective cloud formation through a series of complex processes. Our analysis suggests that a 1 week time lag is required for the effects of these inputs to manifest themselves in streamflow variation in the Mahanadi basin. PA was found to be more important than TPW and TA, perhaps because the pressure anomaly is more directly related to the convective cloud formation.

[72] The importance of large-scale circulation parameters on a local hydrologic phenomenon represents hydroclimatic teleconnection. Among the different large-scale circulation indices, ENSO and EQUINOO are the most important for Indian hydroclimatology. ENSO from past 5th to 7th week and the EQUINOO index from the past 3rd to 6th week were important in explaining streamflow variation in the basin. *Rasmusson and Carpenter* [1983, p. 527] state that "episodes of above normal SST's over the eastern and central equatorial Pacific are associated with a low SOI

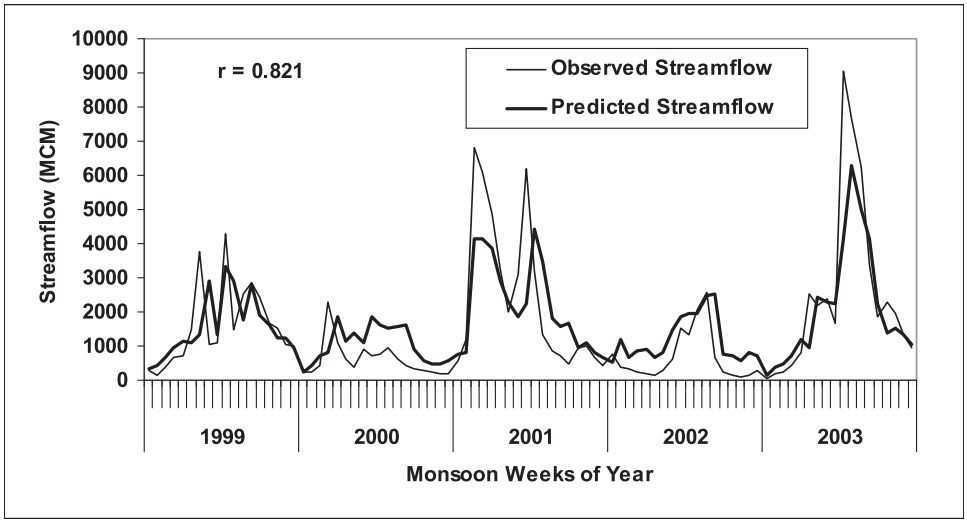

Figure 10. Comparison between observed and predicted streamflow during the testing period for the analysis with 15 selected variables (see Table 2).

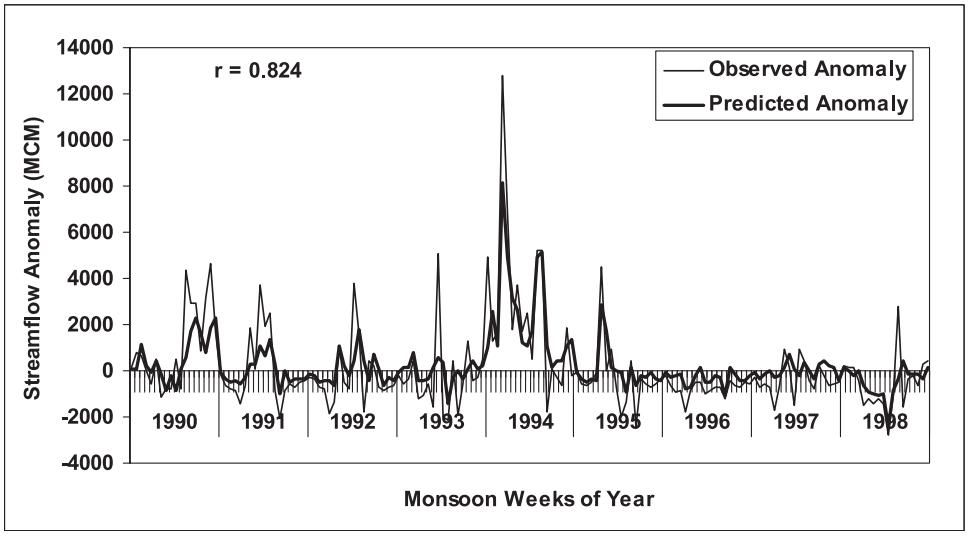

Figure 11. Comparison between observed and predicted streamflow anomalies during the training period for the analysis with 15 selected variables (see Table 2).

(Southern Oscillation Index), i.e., negative pressure anomalies in the southeast Pacific, and positive anomalies over the Indian Ocean region, weaker than normal southwest monsoon over the Arabian Sea, and below normal rainfall over India." Further, *Gadgil et al.* [2003] have shown that the Indian summer monsoon rainfall is associated not only with ENSO but also with EQUINOO. In earlier studies on monthly rainfall, *Maity and Nagesh Kumar* [2006a] have shown that a 2 month lag for ENSO and a 1 month lag for EQUINOO were strong indicators. Similar lags were found to be prominent in the weekly analysis of streamflow here.

[73] Figures 9 and 10 show that extreme values of observed streamflow were not well predicted except in a few instances. This is also reflected in Figures 7 and 8, with all 30 inputs under consideration. However, the trends are well indicated toward the peak flows in most of the cases during the training and testing periods. There could be two reasons. First, inherent uncertainty in both climatic inputs (climatic variables) and output (streamflow) affects the overall performance, particularly for peak flows. GP cannot account for uncertainty in the input and output data. Data uncertainty is present in both the complete set (30 inputs) and the selected set (15 inputs). Errors in data are not necessarily biased toward extreme values. In as much as this work is focused on identifying the important inputs out of many possible potential inputs, the relative importance of an individual input should be less affected by the mismatch of extreme streamflow events. The second reason behind the weaker performance to capture some of the peak flows could be due to nonconsideration of inputs that are important for extreme streamflow events. In other words, the list of inputs considered in the initial model might not be exhaustive. Similar to most if not all modeling approaches, this is a shortcoming of the approach. The existence of

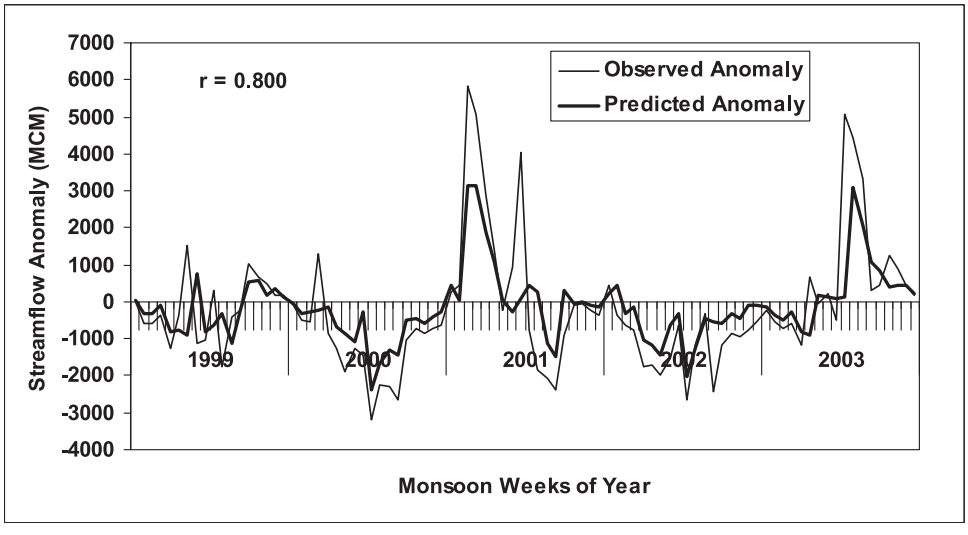

Figure 12. Comparison between observed and predicted streamflow anomalies during the testing period for the analysis with 15 selected variables (see Table 2).

other potential inputs that were not initially incorporated in the model cannot be suggested by the proposed approach. This burden rests with the judgment of the modeler.

6. Conclusions

[74] Importance analysis of various global and local climatic variables that influence weekly streamflow variation is conducted by combining GP results with BIM. While GP is utilized in this study to obtain indices for computing BIM values, other (machine learning) approaches may also be appropriate if a suitable way to measure the contribution of individual variables toward the optimal prediction performance is available. The BIM is a relative measure of importance among different inputs and not an absolute measure. As such, it is not meaningful to prescribe a threshold for acceptability, whereas arranging the inputs in order of their importance is possible. This study highlighted the relative importance of the input variables for basin-scale streamflow prediction in the Mahanadi watershed in India.

[75] It is observed that both large-scale global atmospheric inputs owing to hydroclimatic teleconnection and local inputs feature as important inputs in streamflow prediction. Among the local inputs, OLR and PA are relatively more important than TPW and TA. Among large-scale circulation indices, the ENSO index was important for the past 5–7 weeks, whereas the EQUINOO index was important for the past 3–6 weeks for the basin.

[76] Although temperature plays an important role in rainfall activity at a continental scale, its effect on streamflow was found to be small compared to other local inputs. Other local inputs likely are better measures of moisture and convective activity in the atmosphere leading to precipitation and streamflow. The temperature fluctuates over a wide range on rainy days, thereby diminishing the role of temperature anomalies for streamflow predictions.

[77] The GP approach was able to capture the complex relationship between input variables and streamflow at a weekly temporal resolution. It was found that simultaneous utilization of both large-scale and local influences produced the best prediction performance among all other possibilities, even for such a complex system. This study thus indicates that the effects of both large-scale circulation patterns and local influences are necessary, obviously with different lag effects, to capture the variations in the basin-scale streamflow when rainfall information is either unavailable or unreliable. Apart from the fact that the observed and predicted values correspond well to each other, the trends toward the peak flows are also well indicated in almost all the cases, even though they are not well captured, which can be considered as a future scope of study.

[78] **Acknowledgments.** We acknowledge the help of Govindaraju S. Rao from the School of Civil Engineering, Purdue University, for his help toward the improvement of the overall presentation and English language.

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. Clim.*, *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, doi:10.1029/2001GL013294.

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

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. Plann. Manage.*, *130*(4), 281–289.

Birnbaum, Z. W. (1969), On the importance of different components in a multicomponent system, in *Multivariate Analysis—II*, pp. 581–592, Academic, New York.

Brameier, M., and W. Banzhaf (2001), A comparison of linear genetic programming and neural networks in medical data mining, *IEEE Trans. Evol. Comput.*, *5*(1), 17–26.

Brameier, M., and W. Banzhaf (2007), *Linear Genetic Programming*, 315 pp., Springer, New York.

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.

Chiew, F. H. S., T. C. Piechota, J. A. Dracup, and T. A. McMahon (1998), E1 Niño/Southern Oscillation and Australian rainfall, streamflow and drought: Links and potential for forecasting, *J. Hydrol.*, *204*, 138–149.

Chowdhury, M. R., and N. Ward (2004), Hydro-metrological variability in the greater Ganges-Brahmaputra-Meghna basins, *Int. J. Climatol.*, *24*, 1495–1508.

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*(7), 2133–2141, doi:10.1029/94WR00751.

Elsayed, A. (1996), *Reliability Engineering*, Addison Wesley Longman, Reading, Mass.

Eltahir, E. A. B. (1996), El Niño and the natural variability in the flow of the Nile River, *Water Resour. Res.*, *32*(1), 131–137, doi:10.1029/95WR02968.

Falvey, M., and J. Beavan (2002), The impact of GPS precipitable water assimilation on mesoscale model retrievals of orographic rainfall during SALPEX'96, *Mon. Weather Rev.*, *130*, 2874–2888.

Francone, F. D. (1998), Discipulus owner's manual, fast genetic programming based AIML technology, RML Technol., Littleton, Colo. [Available at http://www.aimlearning.com.]

Gadgil, S., P. N. Vinayachandran, and P. A. Francis (2003), droughts of the indian summer monsoon: Role of clouds over the Indian Ocean, *Curr. Sci.*, *85*(2), 1713–1719.

Gadgil, S., P. N. Vinayachandran, P. A. Francis, and S. Gadgil (2004), Extremes of the Indian Summer monsoon rainfall, ENSO and equatorial Indian Ocean Oscillation, *Geophys. Res. Lett.*, *31*, L12213, doi:10.1029/2004GL019733.

Gairola, R. M., and T. N. Krishnamurti (1992), Rain rates based on SSM/I, OLR and raingauge data sets, *Meteorol. 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, *Clim. Res.*, *1*, 187–197.

Heywood, M. I., and A. N. Zincir-Heywood (2002), Dynamic page based crossover in linear genetic programming, *IEEE Trans. Syst. Man Cybern., Part B*, *32*(3), 380–388.

Hong, Y. S., and R. Bhamidimarri (2003), Evolutionary self-organizing modeling of a municipal wastewater treatment plant, *Water Res.*, *37*, 1199–1212.

Hou, A. Y., D. V. Lendvina, A. M. Da Silva, S. Zhang, J. Joiner, R. M. Atlas, G. J. Huffman, and C. D. Kummerow (2000), Assimilation of SSM/I-derived surface rainfall and total precipitable water for improving the GEOS analysis for climate studies, *Mon. Weather Rev.*, *128*, 509–537.

Jain, S., and U. Lall (2001), Floods in a changing climate: Does the past represent the future?, *Water Resour. Res.*, *37*(12), 3193–3205, doi:10.1029/2001WR000495.

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

Kashid, S. S., S. Ghosh, and S. R. Maity (2010), Streamflow prediction using multi-site rainfall obtained from hydroclimatic teleconnection, *J. Hydrol.*, *395*, 23–38, doi:10.1016/j.jhydrol.2010.10.004.

Koza, J. R. (1992), *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, MIT Press, Cambridge, Mass.

Krishna Kumar, K., B. Rajagopalan, and M. A. Cane (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., 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, doi:10.1029/2000GL011847.

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. Clim.*, *11*, 2898–2909.

Maity, R., and S. S. Kashid (2010), Short-term basin-scale streamflow forecasting using large-scale coupled atmospheric-oceanic circulation and local outgoing longwave radiation, *J. Hydrometeorol.*, *11*(2), 370–387, doi:10.1175/2009JHM1171.1.

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.

Maity, R., and D. Nagesh Kumar (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.*, *7*, 101–107, doi:10.1002/asl.141.

Maity, R., and D. Nagesh Kumar (2008a), Basin-scale streamflow forecasting using the information of large-scale atmospheric circulation phenomena, *Hydrol. Processes*, *22*(5), 643–650, doi:10.1002/hyp.6630.

Maity, R., and D. Nagesh Kumar (2008b), Probabilistic prediction of hydroclimatic variables with nonparametric quantification of uncertainty, *J. Geophys. Res.*, *113*, D14105, doi:10.1029/2008JD009856.

Maity, R., D. Nagesh Kumar, 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*(1), 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, doi:10.1016/j.jhydrol.2008.01.023.

Marcella, M. P., and E. A. B. Eltahir (2008), The hydroclimatology of Kuwait: Explaining variability of rainfall at seasonal and interannual timescales, *J. Hydrometeorol.*, *9*, 1095–1105, doi:10.1175/2008JHM952.1.

Nageswara Rao, G. (1997), Interannual variation of monsoon rainfall in Godavari River basin—Connections with the Southern Oscillation, *J. Clim.*, *11*, 768–771.

Nezlin, N. P., and E. D. Stein (2005), Spatial and temporal patterns of remotely-sensed and field-measured rainfall in southern California, *Remote Sens. Environ.*, *96*, 228–245.

Olsson, J., C. B. Uvo, K. Jinno, A. Kawamura, K. Nishiyama, N. Koreeda, T. Nakashima, and O. Morita (2004), Neural networks for rainfall forecasting by atmospheric downscaling, *J. Hydrol. Eng.*, *9*, 1–12, doi:10.1061/(ASCE)1084-0699(2004)9:1(1).

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*(5), 5341–5350, doi:10.1029/JD093iD05p05341.

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

Ramirez, M. C. V., H. F. Velho, and N. J. Ferreira (2005), Artificial neural network technique for rainfall forecasting applied to the São Paulo region, *J. Hydrol.*, *301*, 146–162, doi:10.1016/j.jhydrol.2004.06.028.

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. Weather 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.

Uvo, C. B., J. Olsson, O. Morita, K. Jinno, A. Kawamura, K. Nishiyama, N. Koreeda, and T. Nakashima (2001), Statistical atmospheric downscaling for rainfall estimation in Kyushu Island, Japan, *Hydrol. Earth Syst. Sci.*, *5*(2), 259–271.

Webster, P., and C. Hoyos (2004), Prediction of monsoon rainfall and river discharge on 15–30 day timescale, *Bull. Am. Meteorol. Soc.*, *85*, 1745–1765, doi:10.1175/BAMS-85-11-1745.

Xiao, Q., X. Zou, and Y. H. Kuo (2000), Incorporating the SSM/I-derived precipitable water and rainfall rate into a numerical model: A case study for the ERICA IOP-4 cyclone, *Mon. Weather Rev.*, *128*, 87–107.

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

S. S. Kashid, Department of Civil Engineering, Walchand Institute of Technology, Solapur 413006, Maharashtra, India.

R. Maity, Department of Civil Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721302, West Bengal, India. (rajib@civil.iitkgp.ernet.in)