Chapter 9

Genetic Programming: A Novel Computing Approach in Modeling Water Flows

Shreenivas N. Londhe and Pradnya R. Dixit

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/48179

1. Introduction

The use of artificial intelligence in day to day life has increased since late 20th century as seen in many home appliances such as microwave oven, washing machine, camcorder etc which can figure out on their own what settings to use to perform their tasks optimally. Such intelligent machines make use of the soft computing techniques which treat human brain as their role model and mimic the ability of the human mind to effectively employ modes of reasoning that are approximate rather than exact. The conventional hard computing techniques require a precisely stated analytical model and often a lot of computational time. Premises and guiding principles of Hard Computing are precision, certainty, and rigor [1]. Many contemporary problems do not lend themselves to precise solutions such as recognition problems (handwriting, speech, objects and images), mobile robot coordination, forecasting, combinatorial problems etc. This is where soft computing techniques score over the conventional hard computing approach. Soft computing differs from conventional (hard) computing in that, unlike hard computing, it is tolerant of imprecision, uncertainty, partial truth, and approximation. The guiding principle of soft computing is to exploit the tolerance for imprecision, uncertainty, partial truth, and approximation to achieve tractability, robustness and low solution cost [1]. The principal constituents, i.e., tools, techniques of Soft Computing (SC) are Fuzzy Logic (FL), Neural Networks (NN), Evolutionary Computation (EC), Machine Learning (ML) and Probabilistic Reasoning (PR). Soft computing many times employs NN, EC, FL etc, in a complementary rather than a competitive way resulting into hybrid techniques like Adaptive Neuro-Fuzzy Interface System (ANFIS).

The application of soft computing techniques in the field of Civil Engineering started since early nineties and since encompassed almost all fields of Civil Engineering namely Structural Engineering, Construction Engineering and Management, Geotechnical

Engineering, Environmental Engineering and lastly Hydraulic Engineering which is the focus of this chapter. The technique of ANN is now well established in the field of Civil Engineering to model various random and complex phenomena. Other techniques such as FL and EL caught attention of many research workers as a complimentary or alternative technique to ANN, particularly after knowing the drawbacks of ANN [2]. The soft computing tool of Genetic Programming which is essentially classified as an Evolutionary Computation (EC) technique has found its foot in the field of Hydraulic Engineering in general and modeling of water flows in particular since last 12 years or so. Modeling of water flows is perhaps the most daunting task ever faced by researchers in the field of Hydraulic Engineering owing to the randomness involved in many natural processes associated with the water flows. In pursuit of achieving more and more accuracy in estimation/forecasting of water related variables the researchers have made of use Genetic Programming for various tasks such as forecasting of runoff with or without rainfall, forecasting of ocean waves, currents, spatial mapping of waves to name a few. The present chapter takes a stalk of the applications of GP to model water flows which will enable the future researchers who want to pursue their research in this field. The chapter is organized as follows. Next section deals with basics of GP. A review of applications of GP in the field of Ocean Engineering is presented in the next section followed by review of applications in the field of hydrology. Few applications in the field of Hydraulics are discussed in the subsequent section. It may be noted that papers published in reputed international journals are only considered for review. Two case studies are presented next which are based on publications of the first author. The concluding remarks and future scope as envisaged by the authors are discussed at the end.

2. The evolutionary computation

The paradigm of evolutionary processes distinguishes between an organism's genotype, which is constructed of genetic material that is inherited from its parent or parents, and the organism's phenotype, which is the coming to full physical presence of the organism in a certain given environment and is represented by a body and its associated collection of characteristics or phenotypic traits. Within this paradigm, there are three main criteria for an evolutionary process to occur as per [3] and they are

- Criterion of Heredity: Offspring are similar to their parents: the genotype copying process maintains a high fidelity.
- Criterion of Variability: Offspring are not exactly the same as their parents: the genotype copying process is not perfect.
- Criterion of Fecundity: Variants leave different numbers of offspring: specific variations have an effect on behavior and behavior has an effect on reproductive success.

The evolutionary techniques can be differentiated into four main streams of Evolutionary Algorithm (EA) development [4] namely Evolution Strategies (ES), Evolutionary Programming (EP), Genetic Algorithms (GA) and Genetic Programming (GP) [5]. However, all evolutionary algorithms share the common property of applying evolutionary processes in the form of selection, mutation and reproduction on a population of individual structures that undergo evolution. The criterion of heredity is assured through the application of a crossover operator, whereas the criterion of variability is maintained through the application of a mutation operator. A selection mechanism then 'favours' the more fit entities so that they reproduce more often, providing the fecundity requirement necessary for an evolutionary process to proceed.

3. Genetic programming:

Like genetic algorithm (GA) the concept of Genetic Programming (GP) follows the principle of 'survival of the fittest' borrowed from the process of evolution occurring in nature. But unlike GA its solution is a computer program or an equation as against a set of numbers in the GA and hence it is convenient to use the same as a regression tool rather than an optimization one like the GA. GP operates on parse trees rather than on bit strings as in a GA, to approximate the equation (in symbolic form) or computer program that best describes how the output relates to the input variables. A good explanation of various concepts related to GP can be found in [5] Koza (1992). GP starts with a population of randomly generated computer programs on which computerized evolution process operates. Then a 'tournament' or competition is conducted by randomly selecting four programs from the population. GP measures how each program performs the user designated task. The two programs that perform the task best 'win' the tournament. GP algorithm then copies the two winner programs and transforms these copies into two new programs via crossover and mutation operators i.e. winners now have the 'children.' These two new child programs are then inserted into the population of programs, replacing the two loser programs from the tournament. Crossover is inspired by the exchange of genetic material occurring in sexual reproduction in biology. The creation of offspring's continues (in an iterative manner) till a specified number of offspring's in a generation are produced and further till another specified number of generations are created. The resulting offspring at the end of all this process (an equation or a computer program) is the solution of the problem. The GP thus transforms one population of individuals into another one in an iterative manner by following the natural genetic operations like reproduction, mutation and cross-over. Figure 1 shows general flowchart of GP as given by [5].

The tree based GP corresponds to the expressions (syntax trees) from a 'functional programming language' [5]. In this type, Functions are located at the inner nodes; while leaves of the tree hold input values and constants. A population of random trees representing the programs is initially constructed and genetic operations are performed on these trees to generate individuals with the help of two distinct sets; the terminal set T and the function set F.

Population: These are the programs initially constructed from the data sets in the form of trees to perform genetic operations using Terminal set and Function set. The function set for a run is comprised of operators to be used in evolving programs eg. addition, subtraction, absolute value, logarithm, square root etc. The terminal set for a run is made up of the values on which the function set operates. There can be four types of terminals namely inputs, constant, temporary variables, conditional flags. The population size is the number of programs in the population to be evolved. Larger population can solve more complicated problem. The maximum size of population depends upon RAM of the computer and length of programs in the population.

4. Genetic operations

Cross over: Two individuals (programs) are chosen as per the fitness called parents. Two random nodes are selected from inside such program (parents) and thereafter the resultant sub-trees are swapped, generating two new programs. The resulting individuals are inserted into the new population. Individuals are increased by 2. The parents may be identical or different. The allowable range of cross over frequency parameter is 0 to 100%

Mutation: One individual is selected as per the fitness. A sub-tree is replaced by another one randomly. The mutant is inserted into the new population. Individuals are increased by 1. The allowable range of mutation frequency parameter is 0 to 100%

Reproduction: The best program is copied as it is as per the fitness criterion and included in the new population. Individuals are increased by 1. Reproduction rate = 100 – mutation rate – (crossover rate * [1 – mutation rate])

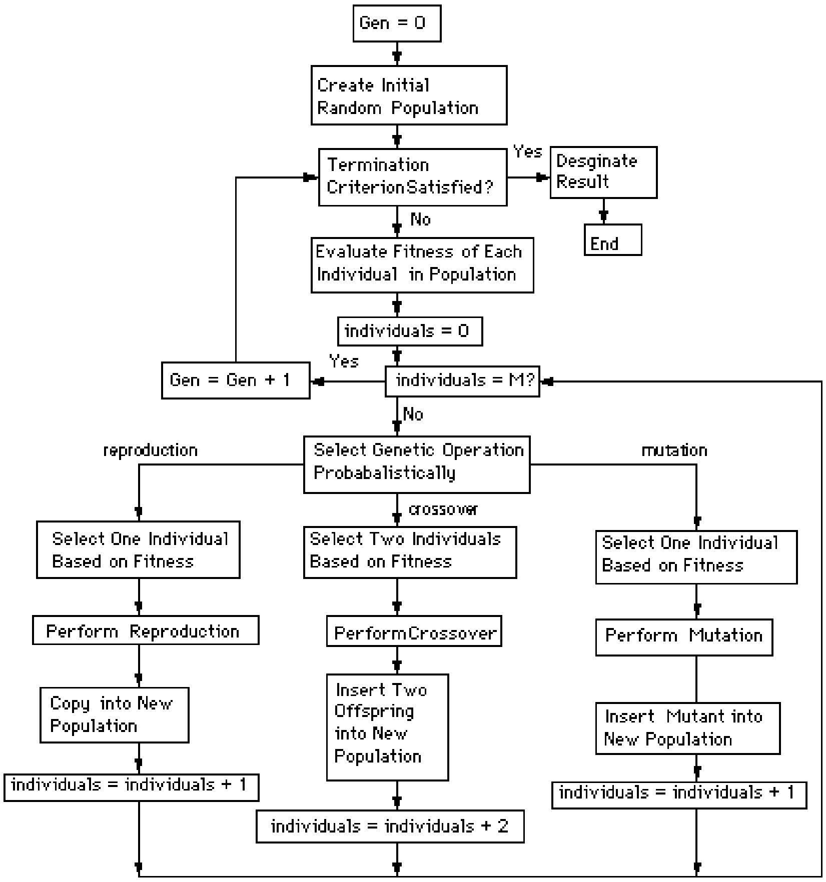
Gen = 0
Create Initial Random Population
Termination CriterionSatisfied?
Yes
Desginate Result
End
No
Evaluate Fitness of Each Individual in Population
individuals = 0
Yes
Gen = Gen + 1
individuals = M?
No
reproduction
Select Genetic Operation Probabalistically
mutation
crossover
Select One Individual Based on Fitness
Select Two Individuals Based on Fitness
Select One Individual Based on Fitness
Perform Reproduction
PerformCrossover
Perform Mutation
Copy into New Population
Insert Two Offspring into New Population
Insert Mutant into New Population
individuals = individuals + 1
individuals = individuals + 2
individuals = individuals + 1

Figure 1. Flowchart of Genetic programming (Ref: [5])

The second variant of GP is Linear genetic Programming (LGP) which uses a specific linear representation of computer programs. The name 'linear' refers to the structure of the (imperative) program representation only and does not stand for functional genetic programs that are restricted to a linear list of nodes only. On the contrary, it usually represents highly nonlinear solutions. Each individual (Program) in LGP is represented by a variable-length sequence of simple C language instructions, which operate on the registers or constants from predefined sets. The function set of the system can be composed of arithmetic operations (+, - , X, /), conditional branches, and function calls (f {x, x^n, sqrt, e^x ,sin, cos, tan, log, ln }). Each function implicitly includes an assignment to a variable which facilitates use of multiple program outputs in LGP. LGP utilizes two-point string cross-over. A segment of random position and random length of an instruction is selected from each parents and exchanged. If one of the resulting children exceeds the maximum length, this cross-over is abandoned and restarted by exchanging equalized segments. An operand or operator of an instruction is changed by mutation into another symbol over the same set. The readers are referred to [7] and [8] for further details.

Gene-Expression Programming (GEP) is an extension of GP, developed by [5]. The genome is encoded as linear chromosomes of fixed length, as in Genetic Algorithm (GA); however, in GEP the genes are then expressed as a phenotype in the form of expression trees. GEP combines the advantages of both its predecessors, GA and GP, and removes their limitations. GEP is a full fledged genotype/phenotype system in which both are dealt with separately, whereas GP is a simple replicator system. As a consequence of this difference, the complete genotype/phenotype GEP system surpasses the older GP system by a factor of 100 to 60,000. In GEP, just like in other evolutionary methods, the process starts with the random generation of an initial population consisting of individual chromosomes of fixed length. The chromosomes may contain one or more than one genes. Each individual chromosome in the initial population is then expressed and its fitness is evaluated using one of the fitness function equations available in the literature. These chromosomes are then selected based on their fitness values using a roulette wheel selection process. Fitter chromosomes have greater chances of selection for passage to the next generation. After selection, these are reproduced with some modifications performed by the genetic operators. In Gene Expression Programming, genetic operators such as mutation, inversion, transposition and recombination are used for these modifications. Mutation is the most efficient genetic operator, and it is sometime used as the only means of modification. The new individuals are then subjected to the same process of modification, and the process continues until the maximum number of generations is reached or the required accuracy is achieved.

5. Why use GP in modeling water flows?

It is a known fact that many variables in the domain of Hydraulic Engineering are of random nature having a complex underlying phenomenon. For example the generation of ocean waves which are primarily functions of wind forcing is a very complex procedure. Forecasting of the ocean waves is an essential prerequisite for many ocean-coastal related activities. Traditionally this is done using numerical models like WAM and SWAN. These models are extremely complex in development and application besides being highly computation-intensive. Further they are more useful for forecasting over a large spatial and temporal domain. The accuracy levels of wave forecasts obtained through such numerical models again leaves scope for exploration of alternative schemes. These numerical models suffer from disadvantages like requirement of exogenous data, complex modeling procedure, rounding off errors and large requirement of computer memory and time and there is no guarantee that the results will be accurate. Particularly when point forecasts were required the researchers therefore used the data driven techniques namely ARMA, ARIMA and since last two decades or so the soft computing technique of Neural Networks. A comprehensive review of applications of ANN in Ocean Engineering is done by [9]. Although wave forecasting models were developed using Artificial Neural Networks by many research workers their was scope for use of another data driven techniques in that the ANN based models generally were unable to forecast extreme events with reasonable accuracy and the accuracy of forecasts decreases with increase in lead time as reported in many research papers. This became an ideal situation for the entry of another soft computing tool of GP which functions in a completely different way than ANN in that it does not involve any transfer function and evolves generations and generations of 'offspring' based on the 'fitness criteria' and genetic operations as explained in the earlier section the researchers thought, may be useful to capture the underlying trends better than ANN technique and can be used as a regressive tool. Same can be said about another important variable in hydraulic engineering "runoff or stream flow".

The rainfall -runoff modeling is very complex procedure and many numerical schemes are available as well as a large number of attempts by ANNs are also been made [2, 10, 11]. Thus Genetic Programming entered in rainfall-runoff modeling. It was also found that GP results were superior to that of M5 Model Trees another data driven modeling technique [12, 13]. Apart from these two variables the use of GP for modeling for many hydraulic engineering processes was found necessary for similar reasons. A review of these applications particularly in Ocean Engineering, Hydrology and Hydraulics (all grouped under Hydraulic Engineering) will be presented in the next three sections.

6. Applications in ocean engineering

As mentioned earlier papers published in reputed international journals are considered in this chapter. Primarily the applications of GP in Ocean Engineering were found for modeling of oceanic parameters like waves, water levels, zero cross wave periods, currents, wind, sediment transport and circular pile scour. Table 1 shows applications of GP in the field of Ocean Engineering listed chronologically followed by their review. This will facilitate the reader to have a glance of the work which would be presented next.

REF. NO.	YEAR	AUTHOR	TITLE OF PAPER	JOURNAL/PUBLICATION
14	2007	Kalra R., Deo M.C.	Genetic Programming to retrieve missing information in wave records along the west coast of India	Applied Ocean Research
25	2007	Singh, A. K., Deo M.C., Sanil Kumar V.	Combined Neural network – genetic programming for sediment transport	Journal of Maritime Engineering, The Institution of Civil Engineers, Issue MAO
16	2007	Charhate S. B., Deo M. C., Sanil Kumar V.	Soft and Hard Computing Approaches for Real Time Prediction of Currents in a Tide Dominated Coastal Area	Journal of Engineering for the Maritime Environment. Proceedings of the Institution of Mechanical Engineers, London, M4
15	2008	Ustoorikar K.S., Deo, M. C.	Filling up Gaps in wave data with Genetic Programming	Marine Structures
18	2008	Jain., P., Deo M. C.	Artificial intelligence tools to forecast ocean waves in real time	The Open Ocean Engineering Journal
22	2008	Charhate, S. B., Deo, M. C., Londhe S. N.	Inverse modeling to derive wind parameters from wave measurements	Applied Ocean Research
17	2008	Gaur, S., and Deo, M. C.	Real time wave forecasting using genetic programming	Ocean Engineering
06	2008	Londhe S. N.	Soft computing approach for real-time estimation of missing wave heights	Ocean Engineering
23	2009	Charhate, S. B., Deo, M. C., Londhe S. N.	Genetic programming for real time prediction of offshore wind	International Journal of Ships and Offshore Structures
26	2009	Guven, A., Azmathulla, H. Md., Zakaria, N.A.	Linear genetic programming for prediction of circular pile scour	Ocean Engineering
24	2009	Daga, M., Deo, M. C.	Alternative data-driven methods to estimate wind from waves by inverse Modeling	Natural Hazards, 49(2), 293-310
08	2009	Guven, A.	Linear genetic programming for time-series modelling of daily flow rate	Journal of Earth Syst. Sci., 118(2), 137-146

19	2010	Kambekar, A. R., Deo, M. C.	Wave simulation and forecasting using wind time history and data driven Methods	Ships and Offshore Structures
20	2010 a	Ghorbani, M. A. , Makarynskyy, O., Shiri, J., Makarynska, D.	Genetic Programming for Sea Level Predictions in an Island Environment	International Journal of Ocean and Climatic systems
21	2010 b	Ghorbani, M. A., Khatibi, R., Aytek, A., Makarynskyy, O., Shiri, J.	Sea water level forecasting using genetic programming and comparing the performance with Artificial Neural Networks	Computers and Geosciences
12	2012	Kambekar, A. R., Deo, M. C.	Wave Prediction Using Genetic Programming And Model Trees	Journal of Coastal Research, Doi: 10.2112/Jcoastres-D-10-00052.1, 28(1), 43-50

Table 1. Applications of GP in Ocean Engineering

One of the earlier applications was done to retrieve missing information in wave records along the west coast of India [14]. Such a need arises many times due to malfunctioning of instrument or drift of wave measuring buoy making it inoperative as a result of which data is not measured and it is lost forever. Filling up the missing significant wave height (Hs) values at a given location based on the same being collected at the nearby station(s) was done using GP. The wave heights were measured at an interval of 3 hours. Data at six locations around Indian coastline was used in this exercise. Out of the total sample size of four years the observations for the initial 25 months were used to evaluate the final or optimum GP program or equation while those for the last 23 months were employed to validate the performance and achieve gap in-filling with different quanta of missing information. It was found that both tree based and linear GP models worked in similar fashion as far as accuracy of estimation was considered. The data was made available by National Institute of Ocean Technology (NIOT) under the National Data Buoy Programme implemented by the Department of Ocean Development, Government of India from January 2000 to December 2003 (www.niot.res.in). The initial parameters selected for a GP run were as follows: initial population size = 500; mutation frequency = 95%; crossover frequency = 50%. The fitness criterion was the mean squared error.

When the similar work was also carried out using ANN it was found that GP produces results that are marginally more satisfactory than ANN. Another exercise was also carried out especially to estimate peaks by calibrating a separate model for high wave data which showed a marginal improvement in prediction of peaks. A similar exercise was carried out by [15], albeit in altogether different area of Gulf of Mexico near the USA coastline. Gaps in hourly significant wave height records at one location were filled by using the significant wave heights at surrounding 3 locations at same time instant and the soft tool of GP and

ANN. In all data spanning over 4 years was used for the study. The exercise was carried out for 4 locations in the Gulf of Mexico. The data can be downloaded from www.ndbc.noaa.gov. The typical value of the population size was 500, number of generations 15 and number of tournaments 90,00,000. The mutation and the cross-over frequency also varied for different testing exercises and it ranged from 20% to 80%. The fitness criterion was the mean squared error between actual observations and corresponding predictions.

The suitability of this approach was also tried for different gap lengths ranging from 1 day to 1 month and it was concluded on the basis of 3 error measures that the accuracy of gap filling decreases with increase in the gap length. The accuracy of the results were also judged by calculating statistical parameters of the wave records without gaps filled and with gaps filled using GP model. When the gap lengths did not exceed 1 or 5 days all the four statistics were faithfully reproduced. Compared to ANN GP produced marginally better results. In both the cases Linear Genetic Programming technique was employed.

In another earlier works of GP current predictions over a time step of twenty minutes, one hour, 3 hours, 6 hours, 12 hours and 24 hours at 2 locations in the tidal dominated area of the Gulf of Khambhat along west coast of India was carried out using two soft techniques of ANN and GP and 2 hard techniques of traditional harmonic analysis and ARIMA [16]. The work involved antecedent values of current only to forecast the current for various lead times at these locations. The fitness function selected was the mean square error, while the initial population size was 500, mutation frequency was 95%, and the crossover frequency was kept at 50%. The authors concluded that the model predictions were better for alongshore currents and small interval of times. For cross shore currents ARIMA performs better than ANN and GP even at longer prediction intervals. In general the three data driven techniques performed better than harmonic analysis. The new technique GP performed at par with ANN if not better. Perhaps the only drawback of the work was that the data (spanning over 7 months) is less than a year indicating that all possible variations in data set were not presented while calibrating the model making it susceptible when it is used at operational level.

Online wave forecasts over lead times of 3, 6, 12 and 24 hours were carried out at two locations in the gulf of Mexico using past values of wave heights (3 in number) and the soft computing technique of GP [17]. The data measured from 1999 to 2004 was available for free download on the web site of National Buoy Centre (http://www.ndbc.noaa.gov). The data belonged to the hourly wave heights measured over a period of 15 years with an extensive testing period of about 5 years which is the most in the papers reported till this time (with ANN as modeling tool). The locations chosen were differing to a large extent in that one was a deep water buoy and the other was a coastal buoy. The work was different from others in one aspect that monthly models were developed instead of routine yearly models. However any peculiar effect of this either good or bad on forecasting accuracy was not evident from the 3 error measures calculated. Though the results of GP were promising (high correlation coefficients for 3 and 6 hr forecast) the forecasting accuracy decreased for longer lead times of 12 hr and 24 hr. It was found that the results of GP were superior to ANN. For GP model the initial population size was 500 while the number of generations was 300. The mutation frequency was 90 percent while the cross over frequency was 50 percent. Values of these control parameters were selected initially and thereafter varied in trials till the best fitness measures were produced. The fitness criterion was the mean squared error between the actual and the predicted value of the significant wave height. Another exercise on real time forecasting of waves for warning times up to 72 hours at three locations along the Indian coastline using alternative techniques of ANN, GP and MT was carried out by [18]. The data was measured from 1998 to 2004 by the national data buoy program (www.niot.res.in). Forecasting waves up to 72hr and that too with reasonable accuracy is itself a specialty of this work. The data had many missing values which were filled by using temporal as well as spatial correlation approaches. Both MT and GP results were competitive with that of the ANN forecasts and hence the choice of a model should depend on the convenience of the user. The selected tools were able to forecast satisfactorily even up to a high lead time of 72 hrs. The authors have rightly stated that this accuracy was possible in the moderate ocean environment around Indian coastline where the target waves were less than around 6 m and 2.5 m for the offshore and coastal stations respectively. The paper does not provide any information about the initial parameters chosen for implementing GP. The significant wave height and average wave period at the current and subsequent 24 hr lead time were predicted from continuous and past 24-hourly measurements of wind speeds and directions as well as two soft computing techniques of GP and MT [19]. The data collected at 8 locations in Arabian Sea and Indian Ocean (www.niot.res.in) was used to develop both hind-casting and forecasting models. Both the methods, GP and MT, performed satisfactorily in the given task of wind wave simulation as reflected in high values of the error statistics of R, R^2, CE and low values of MAE, RMSE and SI. This is noteworthy since MT is not purely non-linear like GP. Although the magnitudes of these statistics did not indicate a significant difference in the relative performance of GP and MT, qualitative scatter diagrams and time histories showed the tendency of MT to better estimate the higher waves. Forecasting at higher lead times were fairly accurate compared to the same at lower ones. In general the performance of wave period was less satisfactory than that of wave height and this can be expected in view of a highly varying nature of wave period values. For details regarding the initial GP parameters involved in calibration readers are referred to the original paper where an exhaustive list of parameters is given. Lately [12], extended their earlier work by forecasting Significant wave height and zero cross wave period over time intervals of 1 to 4 days using the current and previous values of wind velocity and wind direction at 2 locations around the Indian coastline. It was found out that best results were possible when the length of the input sequence matched with that of the output lead time. As observed earlier here also it was found that the accuracy of prediction decreases with increase in lead time. However the results were satisfactory for 4 days ahead predictions also. In general it was observed that results of MT were slightly inferior to that of GP. Separate models were also developed to account for the monsoon (rainfall season in India) which showed a considerable improvement over yearly models. The models calibrated at one location when applied for another nearby locations also shown satisfactory performance provided both sites have spatial homogeneity in terms of openness, long offshore distances and deep water conditions. This work used tree based GP where as earlier mentioned three works used Linear Genetic Programming.

GP was used to forecast sea levels averaged over 12 h and 24 h time intervals for time periods from 12 to 120 h ahead at the Cocos (Keeling) Islands in the Indian Ocean [20]. The model produced high quality predictions over all considered time periods. The presented results demonstrates the suitability of GP for learning the non-linear behavior of sea level variations in terms of the R^2 (with values no lower than 0.968), MSE (with values generally smaller than 431) and MARE (no larger than 1.94%). This differs from earlier applications particularly for wave forecasting in that for forecasting of waves it was difficult to achieve higher order accuracy in terms of r, rmse and other error measures for as far as 24 hour forecast. Perhaps the recurring nature of sea water levels (the deterministic tidal component which is inherent in water level, is the reason behind this high level accuracy. In order to assess the ability of GP model relative to that of the ANN technique, a comparison was performed in terms of the above mentioned statistics. The developed GP model was found to perform better than the used ANNs. In the current work, the linear genetic programming approach was employed. The water level at Hillary's Boat Harbor, Australia was predicted three time steps ahead using time series averaged over 12hr, 24hr, 5 day and 10 day time interval and the soft tool of GP [21]. The results are compared with ANN. Total 12 years of data was used out of which 3 years of data is used for model validation. Tree based GP was used. The results of 12 hr averaged input data were found to be better than 24 hr averaged input data and in general the accuracy of prediction reduced for higher lead times. For both the cases GP results were better than ANN. For 5 day averaged inputs performance of GP was inferior to that of ANN though it improved for 10 day averaged inputs. It may be noted that the input data is averaged over 12hr, 24hr, 5days and 10 days which means there is possibility of loss of information which can be major draw back of this work. For both the above works the hourly sea-level records from a SEA-level Fine Resolutions Acoustic Measuring Equipment (SEA-FRAME) station were used. The information about initial parameters of GP is however not mentioned in both the works.

Estimation of wind speed and wind direction using the significant wave height, zero cross wave period, average wave period and the soft tools of ANN and GP was carried out at 5 locations around Indian coastline [22]. The paper has three folds in that in the first attempt both ANN and GP were tried for estimating the wind speed in which GP was found better and therefore in the second fold GP was only used to determine both wind speed and direction by calibrating the model by splitting of wind vector into two components. Two variants of GP, one based on Tree based approach and the other on Linear Genetic Programming were also tried though the accuracy of estimation for both the approaches was at par. In the third fold a network of wave buoys were formed and wind direction and wind speed at one location was estimated using the same at other locations. This was also done by combining data of all locations and making a regional model. All the attempts yielded highly satisfactory results as far as accuracy of estimation is considered. It was also confirmed that for estimation of only wind speed the non-splitting of wind velocity gives better results. Similarly wind speed and its directions were predicted for intervals of 3hr, 6hr, 9hr, 12hr and 24 hr at locations along the west coast of India using two soft computing techniques of ANN and GP and previous values of the same [23]. It was found that GP rivaled ANN predictions at all the cases and even bettered it particularly for open sea location. The results for prediction of wind speed and wind direction together were better when training of GP and ANN models was done on the basis of splitting of wind vector into two components along orthogonal directions although a separate model for wind speed alone was better (as shown by [22]). In general long interval predictions were less accurate compared to short interval predictions for both the techniques. Data for one location was for about 1.5 years while for the other location it was for 3 years. A discussion on appropriate use of statistical measures to assess the model accuracy was also presented. A similar work was carried out to estimate the wind speed at 5 locations around the Indian coastline using the wave parameters and 3 data driven techniques namely GP (program based- tree type), MT and another data driven tool of Locally weighted projection regression (LWPR) by [24]. All models showed tendency to underestimate higher values in given records. When all of the eight error statistics employed were viewed together, no single method appeared distinctly superior to others, but the use of an average evaluation index EI which they have suggested in this work gave equal weightage to each measure showed that the GP was more acceptable than other methods in carrying out the intended inverse modeling. Separate GP models were developed to estimate higher wind speeds that may be encountered in stormy conditions. At all the locations, these models indicated satisfactory performance of GP although with a fall in accuracy with increase in randomness. For all the above works the data was measured by national data buoy program of India (www.niot.res.in) however no mention is made about the initial parameters chosen for GP implementation.

The estimation of longshore sediment transport rate at an Indian location was carried out using GP and combined GP-ANN models [25]. The data was actually measured by one of the authors in his field study. The inputs were significant wave height, zero cross wave period, breaking wave height, breaking wave angle and surf zone width. The limitation of the work was the amount of data (81) used for training and testing of the models. The choice of control parameters was as follows: initial population size = 500; mutation frequency = 95%; crossover frequency = 50%. The initial trial with GP yielded reasonable results (r = 0.87). However by first training the ANN with same inputs and using the output as input for GP model yielded better results (r = 0.92). Thus the paper shows that combined ANN-GP model is more attractive than single GP model. It may be noted this is a kind of work done in the domain of Ocean Engineering wherein a different parameter (sediment transport rate) is modeled rather than the usual parameters of waves, periods etc. Another different work was carried out by [26], for prediction of scour depth due to ocean/lake waves around a pile/pier in medium dense silt and sand bed using Linear Genetic Programming and Adaptive Neuro-Fuzzy Inference system and measured laboratory data. For initial GP parameters readers are referred to actual paper where in an exhaustive list of parameters is provided. The study was carried out in both dimensional and non-dimensional form in which non-dimensional form yielded better results. The relative importance of input parameters on scour process was also investigated by first using all the influential parameters as inputs and then removing them one by one and observing the results. The drawback of the work is perhaps the small number of data used in model making (total 38 data, 28 of which is used for training the model) which may be impediment in operational use of this model. The results were found to be superior to ANFIS results.

In all the above cases where GP is compared with another data driven technique like ANN, MT or LWPR it was found that GP is superior to all of them in terms of accuracy of results. However it can be said that GP needs to be explored further particularly for prediction of extreme events like water levels, wave heights during hurricanes. A detailed study on effect of variation of GP control parameters like initial population, mutation, crossover percentage etc. on model accuracy is now need of the day. Similarly the critic on other approaches about decreasing forecasting accuracy with increase in the lead time seems to be true for GP as well. This needs more attention if GP is here to stay.

7. Applications in hydrology

Table 2 exhibits the applications of GP in Hydrology chronologically which are reviewed in this paper. The table also indicates that the applications of GP to the field of Hydrology started much earlier as compared to Ocean Engineering.

Genetic Programming is used in Hydrology (science of water) for various purposes such as modeling of phenomena like rainfall-runoff process, evapo-transpiration, flood routing, stage-discharge curve. The GP approach was applied to the flow prediction of the Kirkton catchment in Scotland (U.K.) [27]. The results obtained were compared to those attained using optimally calibrated conceptual models and an ANN. The data sets selected for the modeling process were rainfall, streamflow and Penman open water evaporation. The data used for calibration was of 610 days while that of validation was of 1705 days. The models were developed with preceding values of rainfall, evaporation and stream flow for predicting stream flow one time step ahead. Two conceptual models as well as ANN were employed for developing the stream flow forecasting model. It was observed that the rainfall data was the most influencing factor on the output. All models performed well in terms of forecasting accuracy with GP performing better. The paper does not give any details about the values of the parameters used for calibration of GP model. In another work one day ahead forecasting of runoff knowing the rainfall and runoff of the previous days and the soft computing tool of Linear Genetic Programming was carried out in Lindenborg catchment of Denmark by [28]. The models were developed for forecasting runoff as well as variation of runoff by using previous values of variation of discharge as input as well as previous values of discharge as input along with rainfall information. It was found that it was necessary to include information of discharge rather than variation of discharge. The model predicting discharge gave wrong local peaks in the low regime where as models predicting variation of discharge gave less wrong peaks in the low flow. Both the models had difficulty in predicting high peaks. The models were also developed using ANN. The author concluded that GP is more efficient in peak flow prediction where as ANNs were better in dealing with the noise. The author suggested specialized model for each type of flow to improve the

REF. NO.	YEAR	AUTHOR	TITLE OF PAPER	JOURNAL/PUBLICATION
27	1999	Savic A.D., Walters, G. A., Davidson J.W	A genetic Programming approach to rainfall-runoff modeling	Water Resources Management
28	1999	Drecourt J	Application of Neural Networks and Genetic Programming to Rainfall Runoff Modeling.	Danish Hydraulic Institute (Hydro-Informatics Techonologies - HIT)
29	2001	Whigham, P. A., Crapper, P. F.	Modeling rainfall runoff using Genetic Programming	Mathematical and Computer Modelling,
30	2001	Khu, S. T., Liong, S. U., Babovic, V., Madsen, H., Muttil, N.	Genetic Programming And Its Application In Real-Time Runoff Forecasting	Journal of American Water Resources Association
31	2002	Babovic, V., Keijzer, M.	Rainfall runoff modeling Based on Genetic programming	Nordic Hydrology
32	2007	Sivapragasam,C., Maheswaran, R., Venkatesh, V.	Genetic programming approach for flood routing in natural channels	Hydrological processes
33	2007	Parasuraman, K., Elshorbagy, A., Carey, S. K.	Modelling the dynamics of the evapotranspiration process using genetic Programming	Hydrological Sciences
34	2010	El. Baroudy, I., Elshorbagy, A., Carey, S. K., Giustolisi., O., Savic, D	Comparison of three data-driven techniques in modeling the evapotranspiration process	Journal of Hydroinformatics
13	2010	Londhe, S. N. and Charhate S. B.	Comparison of data driven modeling techniques for river flow forecasting	Hydrological sciences
35	2011	Azmathullah, MD., Ghani, A. AB., Leow, C. S., Chang., C. K., Zakaria, N. A.	Gene-Expression Programming for the Development of a Stage-Discharge Curve of the Pahang River	Water Resource Management

Table 2. Applications of GP in Hydrology

accuracy at peak prediction. He also suggested coupling of black box models with gray models. No specific information is provided about the initial values of GP parameters. The rainfall-runoff relationship in two different catchments was discovered by [29] using GP. The results obtained with a deterministic lumped parameter model, based on the unit hydrograph approach were compared with those obtained using a stochastic machine learning model of GP. For the Welsh catchment in UK, the results between the two models were similar. Since rainfall and runoff were highly correlated, the deterministic assumption underlying the IHACRES model (deterministic) was satisfied. Therefore, IHACREX could achieve a satisfactory correlation between calibration and simulation data. The GP approach which did not require any causal relationships achieved similar results. The behavior of the studied Australian catchment is very different from the Welsh catchment. The runoff ratio was very low (7%), and hence, the a priori assumptions of IHACRES (and other deterministic models) were **a** poor representation of the real world. This was demonstrated by the inability of IHACREJS to use more than one season's data for calibration purposes and only able to use data from a high rainfall period. Since the GP approach did not make any assumptions about the underlying physical processes, calibration periods over more than one season could be used. These led to significantly improved generalizations for the modeled behavior of the catchment. In summary, either approach worked satisfactorily when rainfall and runoff were correlated. However, when this correlation was poor, the CFG-GP had some advantages because it did not assume any underlying relationships. This is particularly important when considering the modeling of environmental problems, where typically the relationships are nonlinear, and are often measured at a scale which does not match with conceptual or deterministic modeling assumptions. Readers are referred to original paper for details of parameters setting for evolving the rainfall-runoff model. In their work of GP in hydrology, [30] first used a simple example of the Bernoulli equation to illustrate how GP symbolically regresses or infers the relationship between the input and output variables. An important conclusion from this study was that non-dimensionalizing the variables prior to symbolic regression process significantly enhance the success of GSR (Genetic Symbolic Regression). GP was then applied to the problem of real-time runoff forecasting for the Orgeval catchment in France. GP functions as an error updating procedure complementing the rainfall-runoff model, MIKE11/ NAM. Ten storm events were used to infer the relationship between the NAM simulated runoff and the corresponding prediction error. That relationship was subsequently used for real-time forecasting of six storm events. The results indicated that the proposed methodology was able to forecast different storm events with great accuracy for different updating intervals. The forecast hydrograph performs well even for a long forecast horizon of up to nine hours. However, it was found that for practical applications in real-time runoff forecasting, the updating interval should be less than or equal to the time of concentration of the catchment. The results were also compared with two known updating methods such as the auto-regression and Kalman filter. Comparisons showed that the proposed scheme, NAM-GSR, is comparable to these methods for real time runoff forecasting. Readers are referred to original paper for details of initial values of various parameters used in calibrating the GP model. The rainfall-runoff models were created on the basis of data alone as well as in combination with conceptual models and Genetic Programming [31]. The study was carried out in Orgeval catchment of France having an area about 104 km^2 using hourly rainfall runoff data of 10 storms for calibration and 6 storms for testing the models. The models were calibrated to forecast the temporal difference between the current and future discharge rather than absolute value of discharge for the lead times of 1 to 12 hours. In fact the paper discusses the phase lag associated with temporal time series forecasting models and removal of it by forecasting the temporal difference. The results were superior to conceptual numerical model. The model was then calibrated using a hybrid method in that the surface runoff value was first forecasted by using a conceptual forecasting model and then using the simulation error and GP to forecast the stream flow. The hybrid models provided a many fold improvement over the raw GP models. The paper in our opinion serves as a basic paper in the field of application of GP in Hydrology and readers may read the paper in original for all details about the GP models developed. The details are not produced here to save the space. Linear Genetic Programming technique was used to predict daily river discharge one day ahead using previous values of the same at Schuylkill River at Berne, PA, USA [8]. Additionally the models were developed using multilayer perceprton as well as Generalized Regression Neural Networks (GRNN). The statistical ARMA method was also used to develop the stream flow forecasting model. The results showed that both LGP and NN techniques predicted the daily time series of discharge with quite good agreement as indicated by high value of coefficient of determination and low values of error measures with the observed data. LGP models generally predicted the maximum and minimum discharge values better than the NN models though LGP results were also far from accurate. The robustness of the developed models was tested by using applied data which was neither used in training or testing and the results were judged using Akaike Information Criterion (AIC). For LGP parameters readers are requested to refer the comprehensive list presented in the paper.

The potential of the GP-based model for flood routing between two river gauging stations on river Walla in USA was explored for single peaked as well as multi-peaked flood hydrographs by [32]. The accuracy of GP models was far superior than modified Muskingum method which is a traditional physics based hydrologic flood routing model which also showed time lag in predictions. The inputs were current and antecedent discharge at upstream station and antecedent discharge at downstream station while the output was current discharge at the downstream station. The LGP was employed for the flood routing exercise. The optimal GP parameters used in this study were: crossover rate, 0.9; mutation rate, 0.5; population size, 200; number of generations, 500; and functional set, i.e. simple arithmetic functions (plus, minus, multiply, divide).

The utility of genetic programming in modeling the eddy-covariance (EC) measured evapo-transpiration flux was investigated by [33]. The performance of the GP technique was compared with artificial neural network and Penman-Monteith model estimates. EC measured evapo-transpiration fluxes from two distinct case-studies with different climatic and topographic conditions were considered for the analysis and latent heat is modeled as a function of net radiation, ground temperature, air temperature, wind speed and relative humidity. Results from the study indicated that both data-driven models (ANN and GP) performed better than the Penman-Monteith method. However, the performance of the GP model is comparable with that of ANN models. One of the important advantages of employing GP to model evapo-transpiration process is that, unlike the ANN model, GP resulted in an explicit model structure that can be easily comprehended and adopted. Another advantage of GP over ANN was found that unlike ANN, GP can evolve its own model structure with relevant inputs reducing the tedious task of identifying optimal input combinations. This work was extended by [34] where in an additional data driven tool of Evolutionary Polynomial Regression was used to model the evapo-transpiration process. Additionally the effect of previous states of input variable (lags) on modeling the EC measured AET (actual evapo-transpiration) is investigated. The evapo-transpiration is estimated using the environmental variables such as net radiation (NR), ground temperature (GT), air temperature (AT), wind speed (WS) and relative humidity (RH). It has been found out that random search and evolutionary-based techniques, such as GP and EPR techniques, do not guarantee consistent performance in all case studies e.g. good and/or bad performance for modelling AET. The authors further stated that this may be due to the practical impossibility of conducting exhaustive search, i.e. searching the entire solution space, to reach the optimal model. The results of ANN, GP and EPR were mostly at par with each other though EPR models were easier to understand. Readers may refer the original papers for above two works for the values of GP parameters.

Recently the stage –discharge relationship for the Pahang River in Malaysia was modeled using Genetic Programming (GP) and Gene Expression Programming (GEP) by [35]. The data was provided by Malaysian Department of Irrigation and Drainage (DID). Gene Expression Programming is an extension of GP. GEP is a full-fledged genotype/phenotype system in which both are dealt with separately, whereas GP is a simple replicator system. Stage and discharge data from 2 years were used to compare the performance of the GP and GEP models against that of the more conventional (stage-rating curve) SRC and (Regression) REG approaches. The GEP model was found to be considerably better than the conventional SRC, REG and GP models. GEP was also relatively more successful than GP, especially in estimating large discharge values during flood events. For details of initial GP parameters the original paper may be referred. The paper elaborates the details of the Gene-expression programming, the new variant of GP.

Like applications in Ocean Engineering it can be said that there is a lot of scope for use of GP in the field of Hydrologic Engineering and more and more applications needs to be tried out.

8. Applications in hydraulics

A few applications of GP in Hydraulic Engineering are also reported in reputed journals which are from open channel hydraulics. Various GP models were developed by [36] to predict velocities across a compound channel with vegetated floodplains. The velocity data was collected in a laboratory flume with steady flow and deep channel and relatively shallow vegetated floodplain on either side. The GP model was developed with all 12 variables in dimensional form depicted accurate results though the evolved equation was complex. The GP models were developed with dimensionless variables and separate for main channel and floodplain. Both the velocity prediction on flood plain and main channels showed good correlations with measured values. However the resulting expressions were complex. A dimensionally aware GP was then used to predict the velocity separately in main channel and flood plains. The performance of the symbolic expressions induced by the dimensionless GP for the floodplain and main channel was marginally better than those for the dimensionally aware GP. However, the expressions were more complex and not particularly useful for knowledge induction. The dimensionally aware GP was shown to hold more scientific information, as units of measurement were included, although it was also shown to be open ended in that it does not strictly adhere to the dimensional analysis framework, thereby allowing improved goodness-of-fit whilst yielding on goodness-of-dimension. The paper provides no information about the initial values of GP parameters used in evolving the GP model. GP was applied to the determination of the Chezy's roughness coefficient for corrugated channels in wake-interference flow, i.e. hyper-turbulent flow by [37]. The GP models were calibrated using the experimental data devised by carrying out experiments for 3 plastic corrugated pipes with variations of discharge and slope. GP quite easily and quickly supplied at least two good formulae that fit the experimental data better and are more parsimonious than the monomial formula (mathematical). Moreover, GP has supplied six parsimonious expressions (one or two constants compared to four for the monomial formula) for the Chezy's resistance coefficient, all confirming the dependencies on hydraulic radius, slope and roughness index. It can be said that the two new formulae for the Chezys resistance coefficient, derived from these GP formulae by means of 'mathematical/physical post-refinement', are suitable for explaining the effect of the macro-roughness elements, with respect to the behavior of the rough commercial channels and their traditional expressions for resistance coefficients. The work indicated that this approach, which combines data-mining techniques together with a theoretical understanding, provides very good results. It was also commented that strictly speaking, GP is a data-driven technique, but prior knowledge during the setting up of the evolutionary search and final physical post-refinement of the hypothesis should make it very close to a white box technique, especially when GP is used in scientific discovery problems. The initial model parameters can be found in the original paper. To save space the list is not provided here.

An alternative approach of GP was proposed in the estimation of relative scour depth using field data by [38]. The comparison between the GP model with ANN found that the GP model has good ability of forecasting the scour depth. The discharge intensity and height of fall were used as inputs to estimate scour depth below tail water. The predictive ability of this approach is however clouded by use of very small number of data (total 91 data sets) used for calibration and testing of the model. The values of initial model parameters can be referred from the original paper.

9. Case study I: Soft computing approach for real-time estimation of missing wave heights

The work dealt with application of GP to retrieve the missing/ lost wave data at a particular location using the wave heights at other locations in the region. Six regional networks (with buoys 42001, 42003, 42007, 42036, 42039,42040) were developed in the Gulf Of Mexico (Figure 2) around USA coastline to estimate the wave heights at a location using wave heights at other five locations in the network. The required data from these six buoys was measured by National Data Buoy Center (NDBC, http://www.ndbc.noaa.gov) of National Oceanic and Atmospheric administration of USA (NOAA, http://www.noaa.gov). The common wave data at all the above six locations for the years 2002-2004 was used in the present work. The networks were developed by having one station as target location at a time and remaining five locations as inputs turn by turn. Approximately 70% of the total values were used to calibrate the model and the remaining was kept unseen for testing. While doing this a particular event which occurred during Hurricane Ivan in 2004 at buoy 42040 which involved a Significant Wave Height of 15.96 m was focused for studying the performance of developed models during extreme events. It is to be noted that the exercise was of estimation and not of forecasting for which both the tools did not performed well as noted in the section on applications of GP in Ocean Engineering.

Thus a network was developed with wave buoy 42040 as the target and buoys 42001, 42003, 42007, 42036, 42039 as inputs. Along with 42040 the other locations namely 42003, 42007, 42039 also experienced largest ever wave heights of 11.04, 9.09, 12.05 making the entire event a truly extra ordinary event having a return period of over 5000 years [39]. The initial parameters selected for a GP run were as follows: initial population size 500, mutation frequency 95%, and crossover frequency 50%. The fitness criterion was the mean squared error.

Additionally a three layer Feed Forward Neural Network was also developed for the same buoy network. The results were also compared with a large-scale continuous wave modeling /forecasting systems (NOAA's WAVEWATCH III model) which follows the approach of physics-based model. Though WAVEWATCH III is a continuous running forecasting model it was the only source of information for wave environment at a location and therefore in absence of any reliable observed data, these results were used for comparison. The GP model estimated a wave height of 13.67m as against 15.96 m as compared to 9.05m that of ANN model and 7.82m of WAVEWATCH III, which was an excellent result as far as GP approach is considered. Figure 3 shows the wave plot at 42040 in testing.

From results of all the models developed by both the approaches (ANN & GP), it was observed that all models performed reasonably well in testing as evident by wave height plots, scatter plots along with the correlation coefficient ranging from 0.85 to 0.98, MAE from 0.13 to 0.28, RMSE from 0.20 to 0.45 m and coefficient of efficiency from 0.67 and 0.96. When it was tried to remove 42001 from the network as it is away from the prevailing wind direction by training a separate GP model with 42003, 42007, 42036, and 42039 as 'input buoys' and 42040 as 'target buoy', though the value of correlation coefficient was increased, the peak prediction was not in a fair range of accuracy for extreme event of Hurricane Ivan. Due to better performance of the network with inclusion of buoy 42001 especially for extreme event, buoy 42001 was retained in the network. Also it was found that 42039 was a potential candidate for redeployment in any other suitable position outside the network as

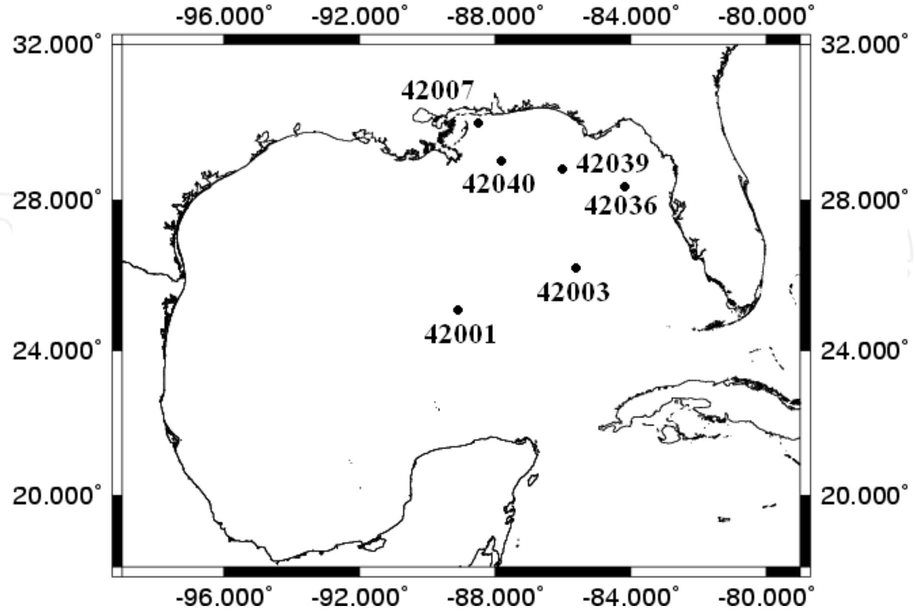

Figure 2. Study area and Buoy Locations (Ref: [6])

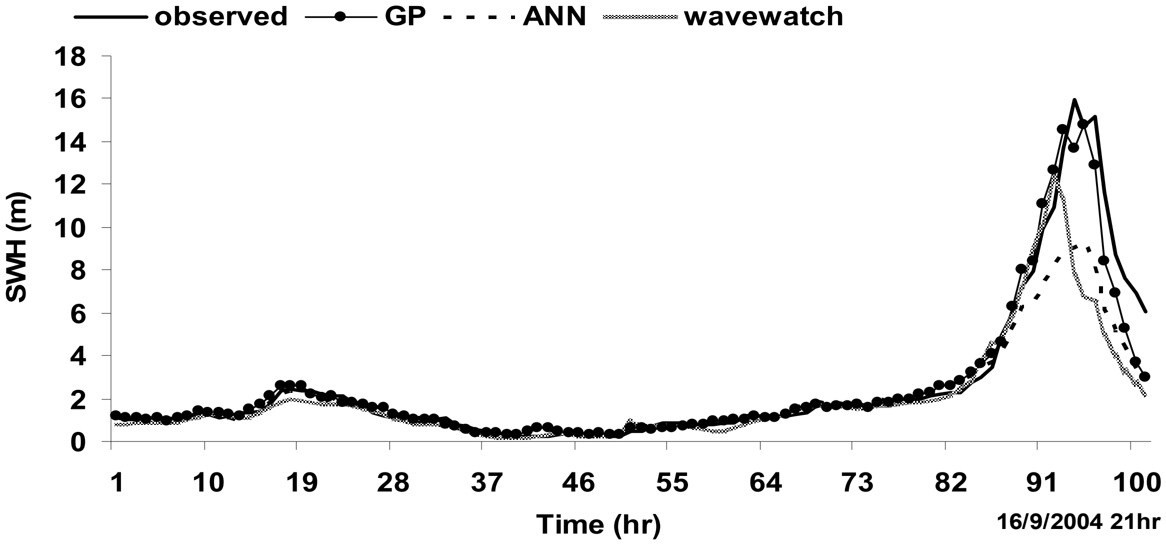

Figure 3. Wave height comparison at 42040 during Hurricane Ivan (Ref: [6])

the buoy network developed for 42039 , provided the wave heights using wave heights at other five locations in the network with the best accuracy achieved between all the networks (r = 0.98). Figure 4(a, b) shows the scatter plots for results of buoy 42039. Table 3 shows results reproduced from [6] giving the details of developed networks along with correlation coefficient between the model estimated and observed values for both GP and ANN models. In general it was shown that GP was superior to other soft tool of ANN and numerical model WAVEWATCH in retrieving the missing wave heights including the extreme events and in redeployment of buoy at other location outside the network.

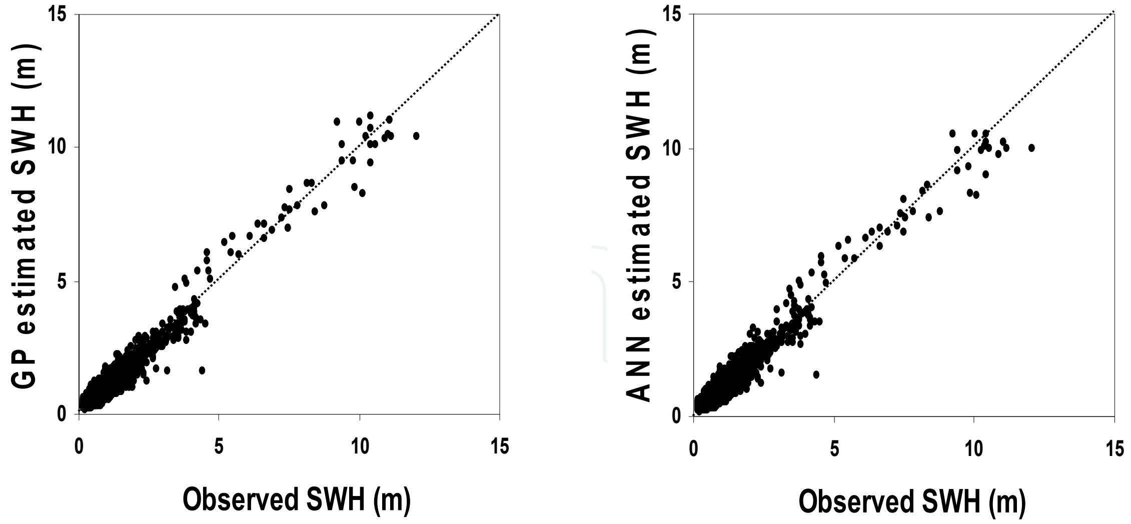

Figure 4. a. Scatter plot for buoy 42039 (GP approach); b. Scatter plot for buoy 42039 (ANN approach) (Ref: [6])

network	Input buoys	Target buoy	r_{ANN}	r_{GP}
BN1	42003, 42007, 42036, 42039, 42040	42001	0.85	0.88
BN2	42001, 42007, 42036, 42039, 42040	42003	0.87	0.91
BN3	42001, 42003, 42036, 42039, 42040	42007	0.90	0.92
BN4	42001, 42003, 42007, 42039, 42040	42036	0.92	0.94
BN5	42001, 42003, 42007, 42036, 42040	42039	0.98	0.98
BN6	42001, 42003, 42007, 42036, 42039	42040	0.94	0.97

Table 3. Results of buoy networks [6]

10. Case study II: Comparison of data-driven modelling techniques for river flow forecasting

In the case study GP was used for prediction of average daily flow values one day in advance at two locations, Rajghat and Mandaleshwar, in the Narmada basin, India using the previous values of measured streamflows at these two locations. The observations of daily average stream flow values at both these stations for the years 1987–1997 were obtained from the Central Water Commission, Narmada Division, Bhopal, India. Considering the variations in daily stream flow values four separate models for the monsoon months of July, August, September and October were prepared along with the one separate but common model for the non monsoon months of November–June. Thus five models were developed in all for each station (total 10 models) to predict discharge at one day in advance. In a view of fair judgment along with GP, ANN and Model trees approach was also employed to develop the models. The number of antecedent discharge values which were used for predicting discharge one day in advance was decided by carrying out the auto-correlation analysis.

The GP models were developed with major fitness function of mean squared error, initial population size of (2048), mutation frequency of (95%) and the cross-over frequency of (53%) with same data division for both ANN and GP models so that their results could be compared. All the developed forecasting models were tested for unseen inputs and their qualitative and quantitative performance was judged by means of correlation coefficient (r) between the observed and forecasted values along with root mean square error (RMSE) and plotting scatter plots between the same. Hydrographs were also plotted to visualize the behavior of the forecasting models particularly for extreme events (peaks).

After examining the results it was observed that for the location of Rajghat in the month of July, ANN model exhibited a reasonable performance in testing with an 'r' value of 0.75 between the observed and forecasted discharges whereas GP model had showed a better 'r' value of 0.78 with better performance for higher values of stream flow, though over-predicted in some instances. The MT model gave a lower 'r' value of 0.7 and prediction of MT model for high stream flows was poor as compared to ANN and GP models. The scatter plot (Fig. 5) between the observed and forecasted discharges confirmed this with a balanced scatter except at the high values of measured stream flows.

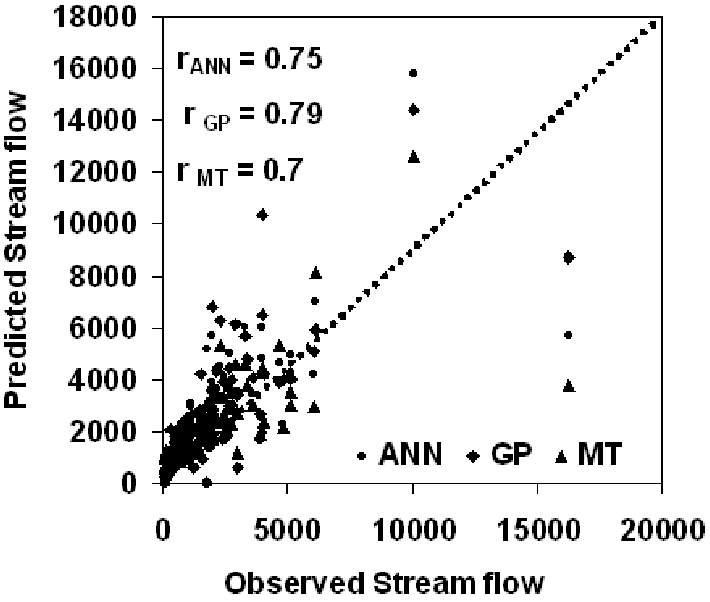

Figure 5. Scatter plot for RajJuly Model

For the months of August and September, models showed similar performance with GP models performing better than their ANN and MT counterparts (r $_{GP}$ = 0.75,r$_{ANN}$ = 0.7, r $_{MT}$ = 0.72 for Raj Aug and r $_{GP}$ = 0.79,r$_{ANN}$ = 0.76, r $_{MT}$ = 0.78 for Raj Sept). For the October model, the predicted discharges in testing were highly in agreement with the observed values for both the models as shown by the discharge hydrograph (Fig. 6). The results were also supported by a high value of correlation coefficient (r = 0.92 for ANN and GP and r = 0.87 for MT) for all the three models in testing.

The Mandaleshwar models behaved in a similar fashion as that of the Rajghat models with correlation coefficients of r > 0.7 for all ANN, GP and MT models. For the month of August the performance of all models was reasonable with r values of 0.74, 0.78 and 0.71 for ANN, GP and MT models respectively. The other monthly models of ANN, GP and MT also performed well, with high correlation coefficients in testing (r > 0.86). It was again observed that GP models work better while predicting extreme events. The maximum observed discharge of 3790 m^3/s was predicted as 1742 m^3/s by the ANN model, 3342 m^3/s by the GP model and 1718 m^3/s by the MT model. Figure 7 shows discharge hydrographs for the ManNov-June models. The RMSE values also showed a similar trend to that of the correlation coefficients.

Thus it was seen that the GP technique outperforms both ANN and MT in almost all the cases in terms of overall accuracy in prediction. The GP approach based on evolutionary principles has a completely different approach to the ANN technique in that it does not involve any transfer function, and evolves generations of "offspring" based on the "fitness criteria" and genetic operations; this seems to capture the underlying trends better than the ANN technique. Thus it can be said that ANN and MT perform almost equally but GP performed better than both of them where prediction accuracy in both normal and extreme events is concerned.

11. Concluding remarks and future scope

Applications of GP for modeling water flows were discussed in the preceding sections of this chapter. It may be noted that every attempt is made to provide readers the details of GP techniques and their parameters employed in each work. However in view of keeping the length of the chapter in stipulated limits sometimes the readers are referred to the original paper. Details about the data are also provided at appropriate locations. Interested readers may further enquire the authors or download the data whenever possible from the web sites to perform the similar exercise. The applications were from three particular areas of water flows namely Ocean Engineering, Hydrology and Hydraulics. It was shown in all the applications for that modeling of natural random processes of complex underlying phenomenon the Genetic Programming can certainly be employed. The results of this technique were found to be superior than other contemporary soft computing techniques. However it was also seen that the tool is not explored to its full capacity by the research community in any of the above fields. The developed GP models also need to be applied at operational level. For this a partnership between the researchers and practitioners is necessary. The GP models can certainly work as supplementary tool if not as replacement techniques. It can be said that the early days of GP modeling are over and the tool needs to be used more judiciously for the problems worthy of its use. Otherwise a stage will be reached where in GP will be used because data is available. It's use is certainly for the phenomena which are difficult to explain and model. However if the technique is to stay here it needs to be explored further for more challenging problems like modeling of infiltration, high flood events, hurricane path, storm surge, tsunami water levels to name a few.

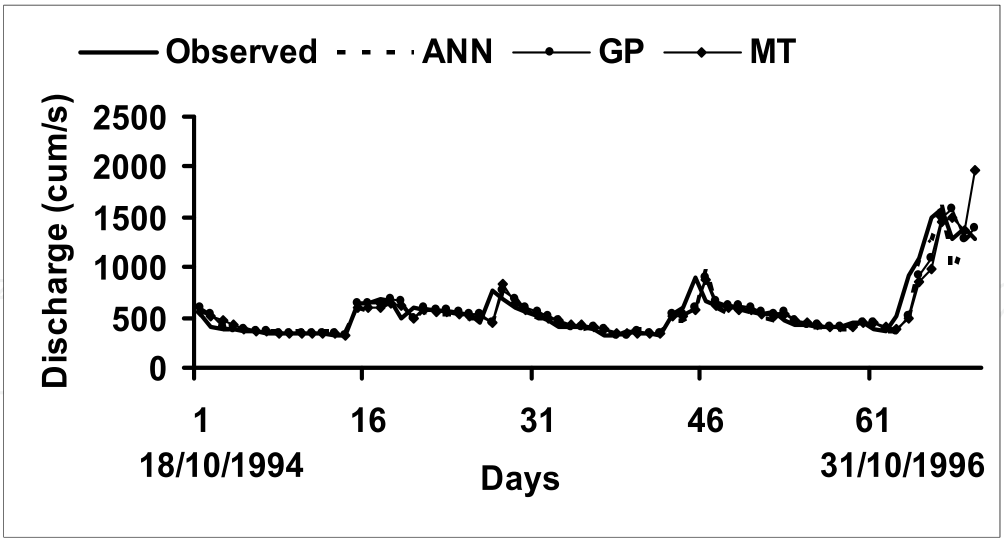

Figure 6. RajOct Model results [13]

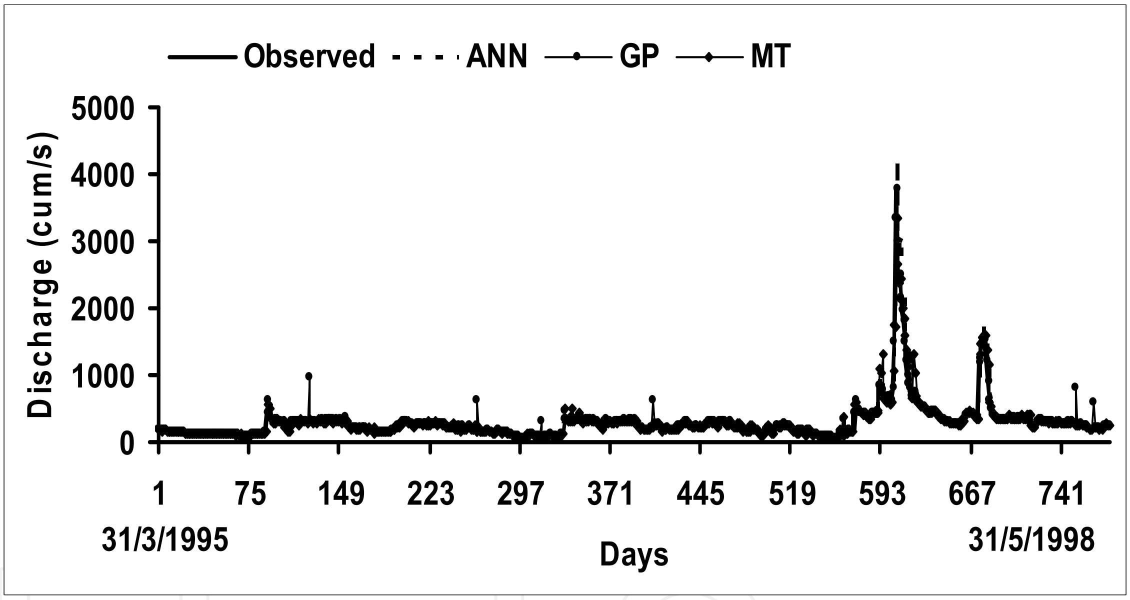

Figure 7. ManNovJune Model results [13]

Author details

Shreenivas N. Londhe and Pradnya R. Dixit
Vishwakarma Institute of Information Technology, Kondhwa (bk), Pune, India
shreel69@yahoo.com, prdxt11@gmail.com

12. References

[1] Zadeh L, (1994) Fuzzy Logic, Neural Networks and Soft Computing. Communications of the ACM 37 (3), 77–84.

[2] The ASCE Task Committee, (2000) Artificial neural networks in hydrology. I: preliminary concepts. J. Hydrol. Engg. ASCE 5(2). 115–123.

[3] Maynard S, (1975) The Theory Of Evolution. Penguin. London.

[4] Babovic V, Keijzer M, (2000) Genetic programming as a model induction engine. Journal of Hydroinformatics. 2(1) pp. 35 – 61

[5] Koza J, (1992) Genetic Programming: On the Programming of Computers by Means of Natural Selection. A Bradford Book. MIT Press.

[6] Londhe S, (2008) Soft computing approach for real-time estimation of missing wave heights. Ocean Engineering. 35. 1080-1089

[7] Brameier M (2004) On linear genetic programming. Ph.D. thesis. University of Dortmund.

[8] Guven A, (2009) Linear genetic programming for time-series modelling of daily flow rate. J. Earth Syst. Sci. 118(2). 137-146

[9] Jain P, Deo M, (2006) Neural networks in ocean engineering. Int. Journal of Ships and Offshore Structures. 1. 25–35.

[10] Maier H, Dandy G, (2000) Neural networks for prediction and forecasting of water resources variables: a review of modelling issues and applications. Environ. Model. Soft. 15. 101–124.

[11] Dawson C, Wilby R, (2001) Hydrological modelling using artificial neural networks. Progr. Phys. Geogr. 25(1). 80–108.

[12] Kambekar A, Deo M, (2012) Wave Prediction Using Genetic Programming And Model Trees. Journal of Coastal Research. Doi: 28(1). 43-50

[13] Londhe S, Charahate S, (2010) Comparison of data-driven modelling techniques for river flow Forecasting. Hydrological Sciences. 55(7). 1163-1173

[14] Kalra R, Deo M, (2007) Genetic Programming to retrieve missing information in wave records along the west coast of India. Applied Ocean Research. 29. 99-111

[15] Ustoorikar K, Deo M, (2008) Filling up Gaps in wave data with Genetic Programming. Marine Structures. 21. 177-195

[16] Charhate S, Deo M, Sanil Kumar V, (2007) Soft and Hard Computing Approaches for Real Time Prediction of Coastal Currents in a Tide Dominated Area. Journal of Engineering for the Maritime Environment. Proceedings of the Institution of Mechanical Engineers, London, M4, 221:147-163

[17] Gaur S, Deo M, (2008) Real time wave forecasting using genetic programming Ocean Engineering. 35. 1166-1175

[18] Jain P, Deo M, (2008) Artificial intelligence tools to forecast ocean waves in real time. The Open Ocean Engineering Journal. 1. 13-21.

[19] Kambekar A, Deo M, (2010) Wave simulation and forecasting using wind time history and data driven Methods. Ships and Offshore Structures. 5(3). 253-266

[20] Ghorbani M, Khatibi R, Aytek A, Makarynskyy O, Shiri J, (2010a) Sea water level forecasting using genetic programming and comparing the performance with Artificial Neural Networks. Computers and Geosciences. 36. 620-627

[21] Ghorbani M, Makarynskyy O, Shiri J, Makarynska D, (2010b) Genetic Programming for Sea Level Predictions in an Island Environment. International Journal of Ocean and Climatic systems. 1(1). pp. 27-35,

[22] Charhate S, Deo M, Londhe S, (2008) Inverse modeling to derive wind parameters from wave measurements. Applied Ocean Research. 30. 120-129

[23] Charhate S, Deo M, Londhe S, (2009) Genetic programming for real time prediction of offshore wind. International Journal of Ships and Offshore Structures. 4(1). 77-88.

[24] Daga, M, Deo M, (2009) Alternative data-driven methods to estimate wind from waves by inverse Modeling. Natural Hazards. 49(2). 293-310

[25] Singh A, Deo M, Sanil Kumar V, (2007) Combined Neural network – genetic programming for sediment transport. Journal of Maritime Engineering, The Institution of Civil Engineers. Issue MAO. 1-7.

[26] Guven A, Azmathulla Md, Zakaria N, (2009) Linear genetic programming for prediction of circular pile scour. Ocean Engineering. 36. 985-991

[27] Savic D, Walters G, Davidson J, (1999) A genetic Programming approach to rainfall-runoff modeling. Water Resources Management. 13. 219-231

[28] Drecourt J, (1999) Application of Neural Networks and Genetic Programming to Rainfall Runoff Modeling. Danish Hydraulic Institute (Hydro-Informatics Techonologies - HIT) June 1999. D2K-0699-1.

[29] Whigham P, Crapper, P, (2001) Modeling rainfall runoff using Genetic Programming. Mathematical and Computer Modelling. 33. 707–721

[30] Khu S, Liong S, Babovic V, Madsen H, Muttil N, (2001) Genetic Programming And Its Application In Real-Time Runoff Forecasting. J of American Water Resources Association. 37(2). 439-450

[31] Babovic V, Keijzer M, (2002) Rainfall runoff modeling Based on Genetic programming. Nordic Hydrology. 33(5). 331-346

[32] Sivapragasam C, Maheswaran R, Venkatesh V, (2007) Genetic programming approach for flood routing in natural channels. Hydrological processes. 22. 623-628

[33] Parasuraman K, Elshorbagy A, Carey K, (2007) Modelling the dynamics of the evapotranspiration process using genetic Programming. Hydrological Sciences. 52(3). 563-578

[34] El Baroudy I, Elshorbagy A, Carey S, Giustolisi O, Savic D, (2010) Comparison of three data-driven techniques in modeling the evapotranspiration process. Journal of Hydroinformatics. 12.4. 365-379

[35] Azmathullah MD, Ghani A, Leow C, Chang C, Zakaria N, (2011) Gene-Expression Programming for the Development of a Stage-Discharge Curve of the Pahang River. Water Resource Management. 25. 2901-2916

[36] Harris E, Babovic V, Falconey R, (2003) Velocity Predictions in Compound Channels with Vegetated Floodplains using Genetic Programming. Int. J. River Basin Management. 1(2). 117-123

[37] Giustolisi O, (2004) Using genetic programming to determine Chezy's resistance coefficient in corrugated channels. Journal of Hydroinformatics. 6.3. 157-173

[38] Azmathullah MD, Ghani A, Zakaria N, Lai S, Chang C, Leow C, (2008) "Genetic Programming to Predict Ski-Jump bucket Spillway Scour. Journal of Hydrodynamics. 20(4), 477-484

[39] Panchang V, Li D, (2006), Large waves in the Gulf of Mexico Caused by Hurricane Ivan. Bulletin of the American Meteorological Society. DOI: 10.1175/BAMS-87-4-481, 481-489.