International Journal of Earth, Energy and Environmental Sciences
ISSN: 2517-942X
Vol:5, No:3, 2011

Exploiting Two Intelligent Models to Predict Water Level: A field study of Urmia lake, Iran

Shahab Kavehkar, Mohammad Ali Ghorbani, Valeriy Khokhlov, Afshin Ashrafzadeh and Sabereh Darbandi

Abstract—Water level forecasting using records of past time series is of importance in water resources engineering and management. For example, water level affects groundwater tables in low-lying coastal areas, as well as hydrological regimes of some coastal rivers. Then, a reliable prediction of sea-level variations is required in coastal engineering and hydrologic studies. During the past two decades, the approaches based on the Genetic Programming (GP) and Artificial Neural Networks (ANN) were developed. In the present study, the GP is used to forecast daily water level variations for a set of time intervals using observed water levels. The measurements from a single tide gauge at Urmia Lake, Northwest Iran, were used to train and validate the GP approach for the period from January 1997 to July 2008. Statistics, the root mean square error and correlation coefficient, are used to verify model by comparing with a corresponding outputs from Artificial Neural Network model. The results show that both these artificial intelligence methodologies are satisfactory and can be considered as alternatives to the conventional harmonic analysis.

Keywords—Water-Level variation, Forecasting, Artificial Neural Networks, Genetic Programming, Comparative analysis.

I. Introduction

WATER level variations are highly sensitive to many environmental forcing, such as lunar and solar gravitational attraction, waves and currents, atmospheric pressure and wind forcing, as well as many other dynamic presumably nonlinear and interconnected physical variables.

Prediction of future water level heights in the coastal environment is of great importance for protection of low-lying regions' residents, for monitoring and prediction of changes in fishery and marine ecosystems. Different methods are used for water level prediction including time series analysis, fuzzy logic, neurofuzzy, genetic programming, artificial neural networks and, recently, chaos theory.

Since the 1990s, timeseries methods employing the Genetic Programming (GP), Artificial Neural Network (ANN) and fuzzy logic methods have become viable, giving rise to the publication of many scientific studies. This paper aims the application of GP and ANN models to forecast sea level timeseries, which are data-driven modeling approaches.

The GP methods are wide-ranging similar to the Genetic Algorithms (GA), first proposed by Koza (1992), as a generalization of GA (Goldberg 1989) [13], [18]. Generally they are robust applications of optimization algorithms and represent one way of mimicking nature. The techniques have the capability for deriving a set of mathematical expressions to describe the relationship between the independent and dependent variables using such operators as mutation, recombination (or crossover) and evolution. As to be elaborated later, these are operated in a population evolving in generations through a definition of fitness and selection criteria, where the subsequent techniques are data-driven. GP techniques are particularly applicable to cases where: (i) the interrelationships among the relevant variables are poorly understood or suspected to be wrong; (ii) finding the size and shape of the ultimate solution is itself a major part of the problem; (iii) conventional mathematical analyses are constrained by restrictive assumptions but approximate solutions are acceptable; (iv) small improvements in performance are routinely measured, easily measurable and highly prized; and (v) the amount of data is large e.g. satellite observation data, requiring examination, classification and integration (Banzhaf et al 1998) [5].

assumptions but approximate solutions are acceptable; (iv) small improvements in performance are routinely measured, easily measurable and highly prized; and (v) the amount of data is large e.g. satellite observation data, requiring examination, classification and integration (Banzhaf et al 1998) [5].

The instantaneous measurements and averaged values of sea level are available in the time and/or space under the influence of variable tides, water temperature etc (e.g. Makarynskyy et al. 2004). They compose time series, and some of the techniques used for their analysis are reviewed below.

Zaldivar et al (1998) used chaos theory techniques for the detection of high water levels in Venice, Italy. Based on their study, non-linear approaches proved capable of simulating dynamic normal trend of water level. Livinia et al (2003) applied stochastic models to estimate the fluctuations of river discharges. Rahmstorf (2007) used a semi-empirical approach to study sea level fluctuations based on earth temperature changes.

Khu et al (2001) applied the GP to real-time runoff forecasting for the Orgeval catchment in France and compared the findings with observed and calculated values using other methods such as the Kalman filter. Their results indicated an acceptable accuracy for the GP. Also, Drecourt (1999) Babovic and Keijzer (2002), Muttil and Liong (2001), Liong et al (2003), and Aytek and Alp (2008) applied the GP for rainfall-runoff modeling. Giustolisi (2004) determined Chezy resistance coefficient using the GP. Borelli et al (2006) introduced an approach based on the GP for extracting the trend in noisy data series. Klara and Deo (2007) applied the GP for filling missing data in wave records along the west coast of India.

Sh. Kavehkar and M. A. Ghorbani and S. Darbandi are with Water Engineering Department, Faculty of Agriculture, University of Tabriz, Tabriz, Iran. Tel: 0098-411-3392786 Fax: 0098-411-3345332, (e-mail: shahab_kvk66@yahoo.com, cusp2004@yahoo.com).

V. Khokhlov are with Hydrometeorological Institute, Odessa State Environmental University, Odessa, Ukraine. (e-mail: vkhokhlov@ukr.net).

A. Ashrafzadeh are with Water Engineering Department, Faculty of Agriculture, University of Guilan, Rasht, Iran. (e-mail: ashrafzadeh@guilan.ac.ir).

Sheta and Mahmoud (2001) forecasted the Nile river flow in the Northern Sudan using the GP. Aytek and Kisi (2008) applied the GP for modeling suspended sediment in streams, concluding that the GP would improve over the conventional rating curves and multi-linear regression techniques and this approach would provide a useful tool in solving specific problems in water resources engineering. Ustoorikar and Deo (2008) used the GP for filling up gaps between data of wave heights. Gaur and Deo (2008) applied the GP for real-time wave forecasting. Ghorbani et al (2010) applied the GP for modeling sea level at the Hillary Boat Harbour and compared the finding with observed and calculated values using artificial neural networks. Their results indicated an acceptable accuracy for the GP.

ANNs can approximate any non-linear mathematical time-series, so the prediction of water level would be achieved with an acceptable accuracy by using ANNs (Hornik, 1993). They have been used extensively for predicting water level fluctuations. Coulibaly et al (2001) used an ANN model for predicting groundwater table fluctuations. Makarynskyy et al (2004) used ANNs for forecasting sea level variations in Hillarys Harbour, Australia. Alvisi et al (2006) predicted water levels using ANNs and fuzzy logic and found that the precision of the ANN over fuzzy logic would be higher whenever more reliable input data are used. Modeling the relationship between water surface and discharge has also been studied using ANNs by Bhattacharya and Solomatine (2005). Chang and Lin (2006) studied multi-point tidal water-level prediction for sites with tidal characteristics similar to a reference site. They expressed tide generating functions in terms of a number of parameters based on essential physical concepts of tidal propagation and tide-generating forces. Using the ANN, they derived the parameters for the various sites in terms of those of a reference site with the trained ANN model. Comparing results with those from a global ocean tidal model, they concluded that their model is applicable if there is a similarity in tide types between the reference site and the application sites, but the applicability reduces as the bathymetric variations become complex [1-26].

II. Materials and Methods

A. Study Area and Data

In this study water level data were obtained from the Local Water Organization of Tabriz, Iran. Figure 1 shows the Urmia lake located at latitude 40.35° North and longitude 13.44° East. Daily sea-water level measurements from January 1997 to July 2008 were used for training and validating of GP and ANN models.

The recorded values range from 1272.55 m (Jul, 2008) to 1277.77 m (June, 1997) with respect to above sea level. However, in a normal year the range of level fluctuations does not exceed 87 cm. The initial timeseries data of water levels were obtained at a daily interval. Table 1 presents some of the important statistics for the time series used and Figure 2 shows the variations of daily data.

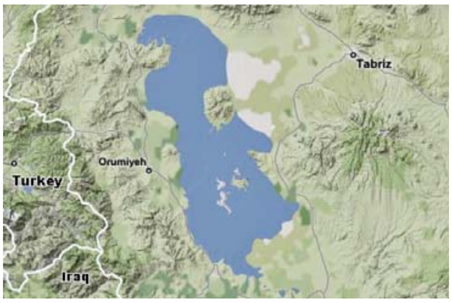


Fig. 1. Location of the Site at Urmia lake

TABLE I
STATISTICS OF DAILY SEA LEVEL DATA FROM URMIA LAKE.

Statistic	Daily sea level (m)
Number of data	4207
Mean	1274.511
Standard Deviation	1.569
Maximum value	1277.77
Minimum value	1272.55
Coefficient of variation	0.00123
Skewness	0.769

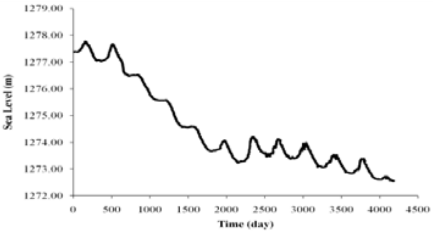


Fig. 2. Daily water level time series at the Urmia Lake

Khatibi (2005) argued that models embody a series of assumptions but after the inception of a project work, the assumptions within the models are progressively challenged, as much as possible by a modelling procedure evolving through the following phases: building raw models, calibration, validation, test controlling for real-time forecasting problems and application. However, in timeseries analysis, it is possible to combine some of these activities after the inception of a project, as follows [16].

Phase 1: review of the data to be used in timeseries and if applicable, identify possible discontinuities in both independent and dependent variables through chaos and catastrophe theory; select the appropriate software application for the timeseries analysis; divide the data into blocks of training data, validation data and application data; and prepare datasets according to the software applications.

Phase 2: implement the timeseries analysis as per selected modelling application; set the parameters appropriate to the selected method of timeseries analysis and the software application; and produce the results.

Phase 3: post-process the results in relation to training, validation and application; and if applicable, carry out some appropriate sensitivity tests. Phase 2 depends on the choice of the timeseries analysis technique. For the GP, the primary objective is to identify a relationship between the independent and dependent variables and for the ANN, the parsimony of the hidden layers of neurons is of critical importance, as discussed below.

B. Genetic programming

The GP is similar to Genetic Algorithm (GA) but employs a parse tree structure for the search of its solutions, whereas the GA employs bite strips. The technique is truly a bottom up process, as there is no assumption made on the structure of the relationship between the independent and dependent variables but an appropriate relationship is identified for any given timeseries. The relationship can be logical statements or it is normally a mathematical expression, which may be in some familiar mathematical format or it may assemble a mathematical functions in a completely unfamiliar format. The GP implementation of relationships has two components:
(i) a parse tree, which is a functional set of basic operators such as $\{+,-,*,' \sqrt{}, \log, a\log, \sin, a\sin, \exp, \cdots\}$ emulating the role of RNA; and (ii) the actual components of the functions and their parameters (referred to as the terminal set), which emulates the role of proteins or chromosomes in biological systems. When these two components work hand-in-hand, only then efficient emulation of evolutionary processes become possible.

The relationship between the independent and dependent variables are often referred to as the model, the program, or the solution but whatever the terminology, the identified relationship in a particular GP modeling is continually evolving and never fixed. As the population evolves from one generation to another, new models replace the old ones by having demonstrably better performance. The evolution starts from an initially selected random population of models, where the fitness value of each model is evaluated using the values of the independent and dependent variables. There are various selection methods and include (i) ranking, in which individual models are ranked and selected according to their fitness values; (ii) selection by tournament, in which the population is regarded as a gene pool of models and a certain number of models are picked up randomly and are then compared according to their fitness; a set number of winners are picked based on their fitness values.

Applying operators like crossover and mutation to the winners, children or offspring are produced, in which crossovers are responsible for maintaining identical features from one generation to another but mutation causes a random change in the parse tree, although data mutation is also possible. This completes the operations at the initial generation and the process is repeated until the termination. There are now various software applications for implementing GP models and Figure 3 presents a typical implementation procedure.

In this study the GP was used for predicting the water level fluctuation. The mathematical form of such a relation can be shown as below:

$$H_{t+\delta\Delta_t} = f(H_t, H_{t-\Delta_t}, \cdots, H_{t-\omega\Delta_t}) \quad (1)$$

In which, H is the height of water level with respect to a refrence point (m), $\delta(\delta = 0, 1, 2, 3, \cdots, \omega)$ describes the time step (Δ_t) used for the forecast water level. The study was carried out by using GeneXpro software application (Ghorbani et al ,2010).

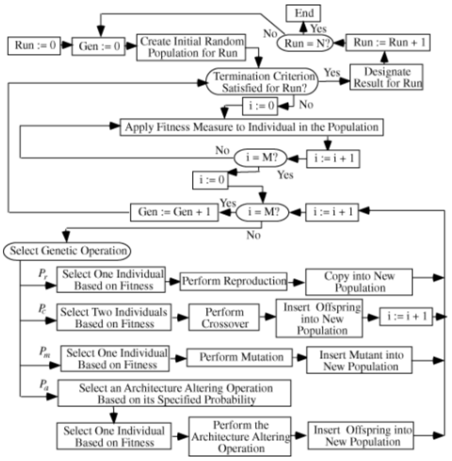

End
Yes
No
Run := 0
Gen := 0
Create Initial Random Population for Run
Run = N?
Run := Run + 1
Termination Criterion Satisfied for Run?
Yes
Designate Result for Run
i := 0
No
Apply Fitness Measure to Individual in the Population
No
i = M?
i := i + 1
i := 0
Yes
Yes
Gen := Gen + 1
i = M?
i := i + 1
No
Select Genetic Operation
P_r
Select One Individual Based on Fitness
Perform Reproduction
Copy into New Population
P_c
Select Two Individuals Based on Fitness
Perform Crossover
Insert Offspring into New Population
i := i + 1
P_m
Select One Individual Based on Fitness
Perform Mutation
Insert Mutant into New Population
P_a
Select an Architecture Altering Operation Based on its Specified Probability
Select One Individual Based on Fitness
Perform the Architecture Altering Operation
Insert Offspring into New Population

Fig. 3. Flowchart of Genetic Programming (Koza, www.genetic-programming.com)

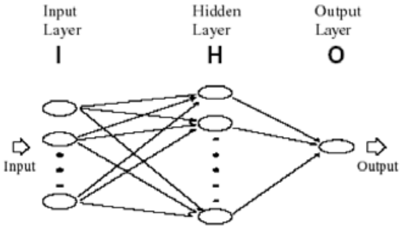


Fig. 4. Neuron Layout of Artificial Neural Networks (ANN)

C. Artificial Neural Networks

ANNs are parallel information processing system. A neural network consists of a set of neurons or nodes arranged in layers and in the case that weighted inputs are used, these nodes provide suitable inputs by conversion functions. Any layer consists of pre-designated neurons and each neural network includes one or more of these interconnected layers. Figure 4 represents a three layered structure that consists of one input layer, I, one hidden layer, H, and one output layer, O. The operation process of these networks is so that the input layer accepts the data and intermediate layer processes them and finally the output layer displays the resultant outputs of model application. During the modeling stage, coefficients related to present errors in nodes are corrected through comparing the model outputs with recorded input data. Further information on ANNs can be found in e.g. Haykin (1999)[15].

The study was carried out by using Qnet software application.

III. RESULTS

The trained and validated model of the data specified for the selected site was used to forecast daily water levels. The performance of the GP and ANN was measured in terms of the correlation coefficient (R) and Root Mean Square Error (RMSE), expressions for which are presented below:

$$R = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} \quad (2)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - x_i)^2}{N}} \quad (3)$$

Where, is the value observed at the ith time step, is the corresponding simulated value, N is number of time steps, is the mean of observational values and is the mean value of the simulations.

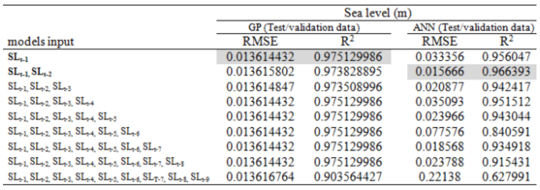
Both the GP and the ANN models were implemented using the recorded data at Urmia lake, Iran. The record covers the years from 1997 to 2008; it was divided into the period from 1997 to 2006 to train model and from 2007 to 2009 to verify it. GeneXpro software was used to implement the GP model with the initial parameters shown in TableII .

TABLE II
CHARACTERISTICS OF EMPLOYED GP.

Genration Without Improvement	150
Generation Since Start	200
Maximum Numbers Of Runs	150
Maximum Program Size	256
Population Size	200
Mutation Rate	90%
Crossover Rate	20%

It is noted that two sets of results are now available: Set 1 produced by the GP approach and Set 2 produced with the ANNs. Figures 5-6 show the recorded and simulated values and their scatter plot.

TABLE III
RMSE AND R2 FOR GENETIC PROGRAMMING AND ARTIFICIAL NEURAL NETWORK MODELS



	Sea level (m)			
	GP (Test/validation data)		ANN (Test/validation data)	
models input	RMSE	R^2	RMSE	R^2
SL_{t-1}	0.013614432	0.975129986	0.033356	0.956047
SL_{t-1}, SL_{t-2}	0.013615802	0.973828895	0.015666	0.966393
SL_{t-1}, SL_{t-2}, SL_{t-3}	0.013614847	0.973508996	0.020877	0.942417
SL_{t-1}, SL_{t-2}, SL_{t-3}, SL_{t-4}	0.013614432	0.975129986	0.035093	0.951512
SL_{t-1}, SL_{t-2}, SL_{t-3}, SL_{t-4}, SL_{t-5}	0.013614432	0.975129986	0.023966	0.943044
SL_{t-1}, SL_{t-2}, SL_{t-3}, SL_{t-4}, SL_{t-5}, SL_{t-6}	0.013614432	0.975129986	0.077576	0.840591
SL_{t-1}, SL_{t-2}, SL_{t-3}, SL_{t-4}, SL_{t-5}, SL_{t-6}, SL_{t-7}	0.013614432	0.975129986	0.018568	0.934918
SL_{t-1}, SL_{t-2}, SL_{t-3}, SL_{t-4}, SL_{t-5}, SL_{t-6}, SL_{t-7}, SL_{t-8}	0.013614432	0.975129986	0.023788	0.915431
SL_{t-1}, SL_{t-2}, SL_{t-3}, SL_{t-4}, SL_{t-5}, SL_{t-6}, SL_{T-7}, SL_{t-8}, SL_{t-9}	0.013616764	0.903564427	0.22138	0.627991

Let SLt represent the sea level at time t. In the present study, the following combinations of input data of sea level were evaluated:

1. SLt-1 ;
2. SLt-1 and SLt-2 ;
3. SLt-1, SLt-2 and SLt-3 ;
4. SLt-1, SLt-2, SLt-3 and SLt-4 ;
5. SLt-1, SLt-2, SLt-3, SLt-4 and SLt-5 ;
6. SLt-1, SLt-2, SLt-3, SLt-4, SLt-5 and SLt-6 ;
7. SLt-1, SLt-2, SLt-3, SLt-4, SLt-5, SLt-6 and SLt-7 ;
8. SLt-1, SLt-2, SLt-3, SLt-4, SLt-5, SLt-6, SLt-7 and SLt-8 ;
9. SLt-1, SLt-2, SLt-3, SLt-4, SLt-5, SLt-6, SLt-7, SLt-8 and SLt-9

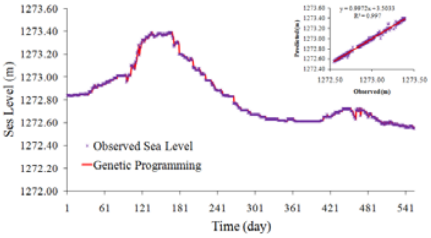


Fig. 5. Comparison between time series plots of predicted and observed values; and Scatter plot of observed and predicted values, (Jan 2007 Jul 2008)

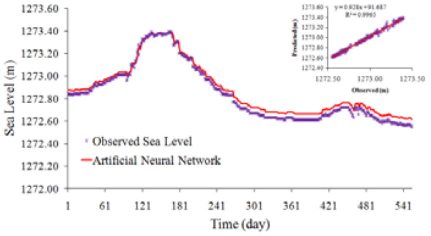


Fig. 6. Comparison between time series plots of predicted and observed values; and Scatter plot of observed and predicted values, (Jan 2007 Jul 2008)

The output layer had one neuron for current sea level SLt.

Table 4 shows the results of each model .It can be seen that the Genetic Programming in general have smaller RMSE values than the Artificial Neural Network model for validation period and the correlation coefficient is high for the Genetic Programming than ANN model.

TABLE IV
STATISTICAL ANALYSIS OF FORECASTED VALUES WITH GP AND ANN METHODS

Model	RSME(m)	R^2	result
Geneting Programing	0.0136	0.997	Genetinc programing is better than Artificial Netural Netwok
Artificial Netural Netwok	0.0156	0.996	

IV. Conclusion

The GP and ANN are used for forecasting water level variations in Urmia lake, Iran. This study uses GeneXpro and the results are compared with those from the ANNs by Qnet software. The mathematical modeling techniques for the analysis of timeseries are diversifying and this paper compared the performance of two such techniques. The GP seems to perform marginally better for most of the cases. The results seem to support the emerging consensus that a single modeling technique is unlikely to render the solution. Instead a set of solutions using different parameters and different simulation techniques is likely to identify the variability of the problem and the solution should be conditioned by the variability.

ACKNOWLEDGMENT

The authors are grateful to the Local Water Organization of Tabriz (Iran) for making available the water-level observations.

REFERENCES

[1] Alvisi, S., Mascellani,G., Franchini, M., Bardossy,A. Water level forecasting through fuzzy logic and artificial neural network approaches. Hydrology and Earth System Science (2006), 10(1), 1-17.
[2] Aytek, A., Alp, M. An application of artificial intelligence for rainfall runoff modeling. Journal of Earth System Science (2008), 117(2),145-155.
[3] Aytek, A., Kisi, O. A genetic programming approach to suspended sediment modeling. Journal of Hydrology (2008), 351, 288-298.
[4] Babovic, V., Keijzer, M. Rainfall runoff modeling based on genetic programming . Nordic Hydrology (2002), 33, 331-343.
[5] Banzhaf W, Nordin P, Keller PE, Francone FD. Genetic Programming, Morgan Kaufmann, San Francisco, CA (1998).
[6] Bhattacharya, B., Solomatine, D.P. Neural networks and M5 model trees in modeling water level-discharge relationship. Neurocomputing (2005), 63, 381-396.
[7] Borelli A, De Falco I, Della CA, Nicodemi M, Trautteur G. Performance of genetic programming to extract the trend in noisy data series. Physica A (2006), 370: 104-108.
[8] Chang H-K and Lin L-C. Multi-point tidal prediction using artificial neural network with tide-generating forces, Coastal Engineering (2006), 53, P.P. 857864.
[9] Coulibaly, P., Anctil, F., Aravena, R., Bobee, B. Artificial neural network modeling of water table depth fluctuation. Water Resources Researches (2001), 37(4), 885-896.
[10] Gaur S, Deo MC. Real-time wave forecasting using genetic programming. Ocean Engineering (2008), 35(11-12):1166-1172.
[11] Ghorbani MA, Khatibi R, Aytek A, Makarynskyy O, Shiri J. Sea Water Level Forecasting Using Genetic Programming and Comparing the Performance with Artificial Neural Networks. Computers & Geosciences (2010), 36:620627.
[12] Giustolisi, O. Using GP to determine Chezzy resistance coefficient in corrugated channels. Journal of Hydroinformatics (2004), 157-173.
[13] Goldberg DE. Genetic algorithms in search, optimization, and machine learning. Addison Wesley, Reading, Mass (1989).
[14] Hornik, K. Some new results on neural network approximation. Neural Networks (1993), 6, 1069-1072.
[15] Haykin, S. Neural networks: a comprehensive foundation, Prentic-Hall, Upper saddle river, New Jersey (1999), 842 PP.
[16] Khatibi, R. Barriers inherent in Flood Forecasting and their Treatments, Chapter 29 of the book: River Basin Management for Flood Risk Mitigation, Ed. D.W. Knight and A.Y. Shamseldin (2006).
[17] Khu, S.T., Liong, S.Y., Babovic, V., Madsen, H., Muttil, N. Genetic programming and its application in real- time runoff forming. Journal of American Water Resources Association (2001), 37(2), 439-451.
[18] Koza JR. Genetic Programming: On the programming of computers by means of Natural Selection. Cambridge, MA: The MIT Press (1992).
[19] Liong, S.Y., Gautam, T.R., Khu, S.T., Babovic, V., Keijzer,M., Muttil, N. Genetic programming: A new paradigm in rainfall runoff modeling. Journal of American Water Resources Association (2002), 38(3), 705-718.
[20] Livinia V, Ashkenazy Y, kinzer Z, Stryging V, Bunde A, HAvlin S. A stochastic model of river discharge fluctuation. Physica A (2003), 330:283-290.
[21] Makarynskyy O, Makarynska D, Kuhn M, Featherstone WE. Predicting sea level variations with artificial neural networks at Hillary Harbour, Western Australia. Estuarine, Coastal and Shelf Science (2004), 61: 351-360.
[22] Muttil, N., Liong, S.Y. Improving runoff forecasting by input variable selection in GP. In: Proceedings of world water congres, ASCE (2001).
[23] Rahmstorf, S. A semi empirical approach to projecting future sea level rise. Science (2007), 315 (5810), 368-370.
[24] Sheta, A.F., Mahmoud, A. Forecasting using genetic programming. Proceedings the 33-rd southeastern symposium on system theory (2001), 343-347.
[25] Ustoorikar K, Deo MC. Filling up gaps in wave data with genetic programming. Marine Structures (2008), 21: 177-195.
[26] Zaldivar, J.M., Strozzi, F., Gutierrez, E., Shepherd, I.M. Early detection of high water at Venice Lagoon using chaos theory techniques. European report 17317. ISPRA: E.C (1998).