ARIMA Models versus Gene Expression Programming In Precipitation Modeling

ALINA BĂRBULESCU and ELENA BĂUTU
Department of Mathematics and Computers Science
Ovidius University
124 Mamaia Blv., Constanța, 900527
ROMANIA
alinadumitriu@yahoo.com http://alina.ilinc.ro/ID_262/index.html
ebautu@univ-ovidius.ro http://csam.univ-ovidius.ro/~ebautu

Abstract: In this paper we present a case study: the application of some conceptually different approaches to the problem of identifying a model for a hydrological time series. The problem is particularly challenging, due to the size of the time series and more importantly, to the many complex phenomena that influence such time series and that reflect in the characteristics of the data. We use well established statistical methods to detect change points in the time series, and we model the subseries obtained by ARIMA, GEP and the adaptive variant and a combination of the two. The models obtained state the efficiency of combining pure statistical tests and methods with heuristic approaches.

Key-Words: Time series modeling, Gene Expression Programming, ARIMA, Statistical analysis, Precipitation

1 Introduction

Time series are ubiquitous in the real world. They are usually generated by dynamical systems and can be encountered in any field of science (e.g. periodic records of the unemployment rate, series of exchange rates, daily temperatures, number of failures of equipment per unit of time, etc.). The problem of modeling time series has been the subject of many research reported in the meteo-hydrological literature [1], [7], [8]. Hydrological records hold important information with respect to the study of weather phenomena or the environment. The importance of finding a well fit model for a hydrological time series cannot be denied, since the main revenue obtained by a successful approach would be a model that explains the past very well and provides informed insight into what will happen in the future.

The methods used for time series modeling can be grouped into two broad classes: classical and modern heuristic methods. Classical approaches include exponential smoothing, autoregressive or threshold methods [16]. The majority of heuristic approaches to the problem use neural networks or evolutionary computation [16]. Many algorithms rely on the assumption of a constant data generating process, which implies that once a model that fits a given set of data (a time series of a given size), the problem is solved, and the model may be used to characterize the future. This kind of approach is flawed, since in the real world, the conditions are permanently changing, and the changes in the environment that produces the given time series trigger changes in the gathered data.

In statistics, the point where a change occurs in the data generating process is called a change point. The problem of identifying such points in a time series is referred to as the change point problem.

In this paper, we follow the three step methodology implied by treating the time series from a change point perspective. Therefore, we try to identify the number of change points (if there exist any), their location, and then to model the distribution of the subseries defined.

Then we use both a classical approach – the ARIMA method, and a heuristic one – using standard GEP and the adaptive Gene Expression Programming variant AdaGEP [4], to model the subseries.

We choose use GEP since it is known that evolutionary computation techniques, in particular genetic programming and its variants, have been used with very good results on real world time series [16].

Since the break tests gave contrasting results, we report the models obtained on the entire time series as well using the two modeling techniques mentioned (ARIMA and GEP). We also report some results obtained by combining an autoregressive model with GEP.

1.1 Related Work

Numerous attempts to solve the change point problem are reported. The works of Pettitt [14] and Buishand [6] are seminal in the literature. They propose tests that permit to find of a change point in the mean of a non-stationary time series. Then the time series may be divided into two stationary time series. The assumption that there is only one change point is an important

shortcoming, so Hubert proposes a procedure that builds on the works of Klemes and Potter and optimally yields a partition of the series in many subseries [11].

Once the change points are identified, the problem of identification of a suitable model is split into the determination of models that describe the subseries delimited by the change points. At this step, both classical and modern time series modeling methods may be used.

Piecewise linear approximation (PLA) algorithms are highly used. In [12] authors report good results obtained in conjunction with a symbolic representation. PLA is used to model segments between feature points in [17]. Its major disadvantage is the approximation of the subseries by linear models. The work presented in [9] is a hybrid approach, that combines metaheuristics with classical statistical methods, namely the autoregressive model (AR). Davis uses a genetic algorithm for change point detection followed by autoregressive modeling of the segments determined.

2 Time series modeling

A time series model for the observed data (x_t) is a specification of the joint distributions of a sequence of random variables (X_t) of which (x_t) is postulated to be a realization.

In what follows we shall denote by *n* use the selection volume.

2.1 ARIMA models

For a description of Box-Jenkins methodology, we recommend [5]. We resume ourselves to only list here the notions used later on in the paper.

Let us consider the operators defined by:

$$B(X_t) = X_{t-1},$$

$$\Phi(B) = 1 - \varphi_1 B - ... - \varphi_p B^p, \varphi_p \neq 0,$$

$$\Theta(B) = 1 - \theta_1 B - ... - \theta_q B^q, \theta_q \neq 0,$$

$$\Delta^d (X_t) = (1 - B)^d X_t.$$

The process (X_t) is said to be an ARIMA(*p*, *d*, *q*) process if $\Phi(B)\Delta^d X_t = \Theta(B)\varepsilon_t$, where the absolute values of the roots of Φ and Θ are greater than 1 and (ε_t) is a white noise.

Particular cases are: ARMA(*p, q*)=ARIMA(*p*, 0, *q*), AR(*p*) = ARIMA(*p,* 0, 0), MA(*q*) = ARIMA(0, *d,* 0).

2.2 Heuristic approach

Apart from the statistical treatment by means of ARIMA models, we use a heuristic approach based on GEP [10], [4], [3] to discover fit models for the time series.

The method used by us dynamically adapts the size of GEP chromosomes by adjusting actually the number of genes that actively participate in the decodification process of GEP chromosomes. In standard GEP the size of the individuals is fixed and all genes participate in the decodification process.

The adaptive mechanism use by us resides on the feedback received from the algorithm by means of each individual's fitness. The fitness is measured in terms of prediction error:

$$error = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}.$$

We report also the ratio of prediction error over standard deviation as a measure of the predictions' quality in a model.

2.2.1 Experimental setup

AdaGEP extension implemented for the `gep` package of the framework ECJ[1] has been utilized for this work The number of genes in the GEP chromosomes was set to 6. This means that each enhanced AdaGEP chromosomes has associated a genemap of 6 bits. The size of the head was 5 symbols, the population size 1000, the stopping criterion used a maximum number of generations of 200. The operator rates were left at the default values provided by the framework. The function set included the arithmetic operators $\{+,-,*,/\}$, and also trigonometric functions $\{\sin, \cos\}$. The individuals are evaluated using the prediction error, such that the algorithm favors individuals with smaller prediction error. The selection scheme used was roulette wheel selection, enhanced with elitist survival of the best 10% of the individuals in each generation onto the next. The genetic algorithm that evolves the genemaps used a mutation rate of 0.001 and a crossover rate of 0.65.

2.3 Data analysis

In order to perform the data analysis the following procedures and statistical tests were used:

1. Q-Q plot or Jarque – Bera test – to verify the normality hypothesis [15];
2. The autocorrelation function [5] – to test the hypothesis that the series is uncorrelated;
3. Buishard [6] and Pettitt [14] tests and Hubert's segmentation procedure [11], to determine the existence of change points (breaks);

[1] ECJ is an open-source evolutionary computation research system developed in Java at George Mason University's Evolutionary Computation Laboratory and available at http://cs.gmu.edu/~eclab/projects/ecj/

4. Bartlett test [2] for homoscedasticity.

The series studied are represented in Figs. 1 and 2.

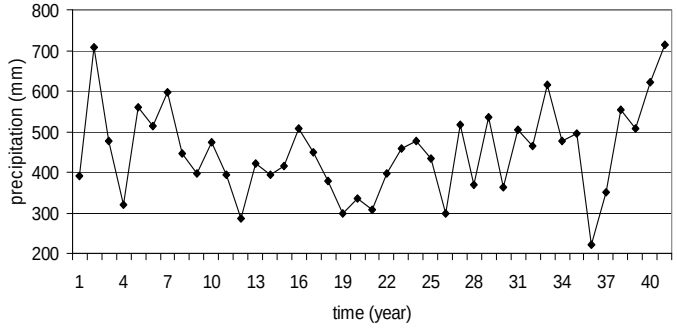

Fig. 1. S_1: The mean annual precipitation (January 1965 - December 2005)

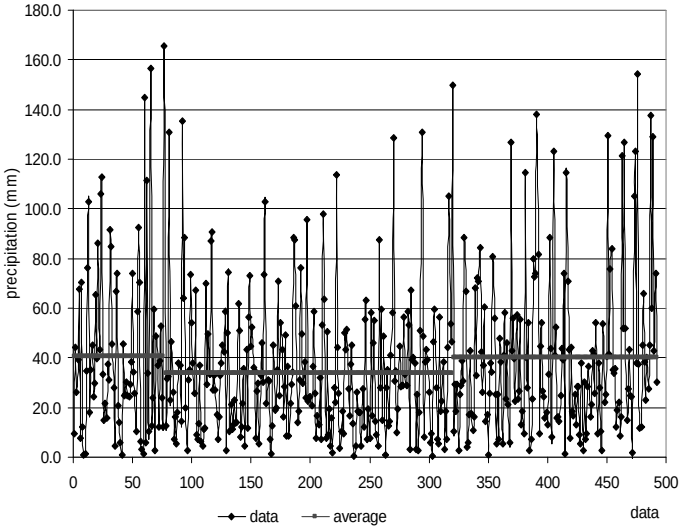

Fig. 2. S_2: The mean monthly precipitation (January 1965 - December 2005)

The tests' results were, respectively:

- For S_1:

The series is normally distributed, independent and homoscedastic [3]

- For S_2:

1. The series doesn't have a normal distribution.

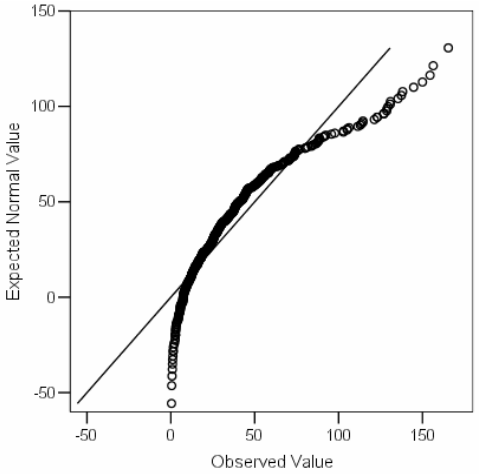

Fig.3. Q - Q plot of S_2

The Q-Q plot diagram (Fig. 3) shows that the observed values are not distributed along the straight line that represents the theoretical normal distribution.

The test Jarque – Bera applied to the series obtained after a Box-Cox transformation,

$$Z_t = \frac{X_t^{\lambda} - 1}{\lambda},$$

with $\lambda = 0.39$, leads us to accept the hypothesis that the series is normally distributed. In addition, the associated histogram (Fig. 4) confirms the normality (the curve draws the chart of theoretical standard Gaussian distribution).

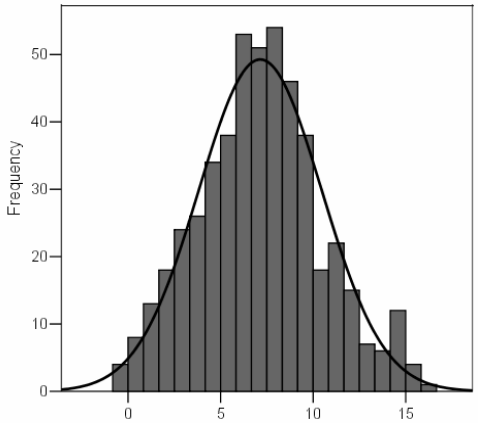

Fig.4. Histogram of the transformed series, (Z_t)

2. The original data series and the transformed one are correlated, since there are values of autocorrelation function (ACF) outside the confidence interval at 95% confidence level. (Fig.5).

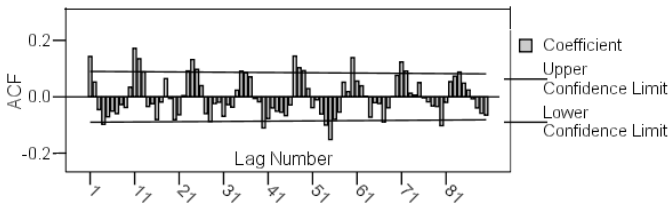

Fig.5. Autocorrelogram of S_2 transformed

3. The results obtained in the break tests are contradictory.

Let the null hypothesis be:

H_0: The time series doesn't have breaks.

Bois' ellipse, associated with Buishard's test is represented in Fig.6. We conclude that H_0 can be accepted at a confidence level of 95%. The Pettitt test leads us to the same conclusion (Fig.7).

H_0 is rejected after the application of Hubert's segmentation procedure. Two break points were determined: in April 1971 and July 1991. The precipitation levels recorded in May 1972 and June 1990 are outliers.

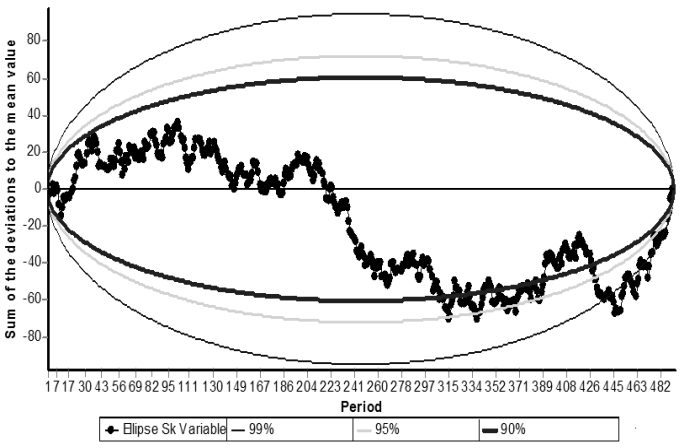

Fig.6. Bois' ellipse

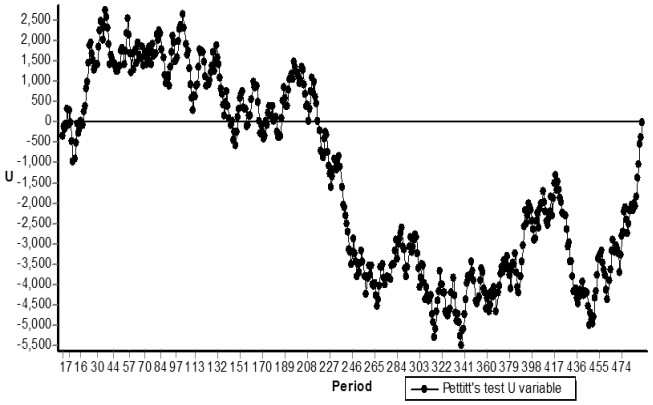

Fig.7. Pettitt test

Given these break-points, the original series S_2 will be split into three subseries, denoted respectively by: S_21 the subseries up to the first break point), S_22 (the subseries consisting of the values between the two break points), and S_23 (the subseries beginning with the second break point up to the end).

The values of some descriptive statistics of these series are given in Table 1.

Table 1. Descriptive statistics

Series	S_2	S_21	S_22	S_23
min	0.4	0.9	0.4	0.7
max	165.4	156.4	135.4	154.5
mean	37.49	40.8	34.12	41.29
median	29.6	32.4	28.35	30.35
variance	950.77	1153.11	685.38	1045.773
std.dev.	30.83	33.96	26.18	32.34

4. Bartlett test was applied dividing S_2 in S_21, S_22, S_23 and also the subseries with the selection volume 164. In both cases, the homoscedasticity hypothesis was rejected.

The change point analysis, based on the cumulated sums, CUSUM, reveals a change point in S_2 (Fig.8)

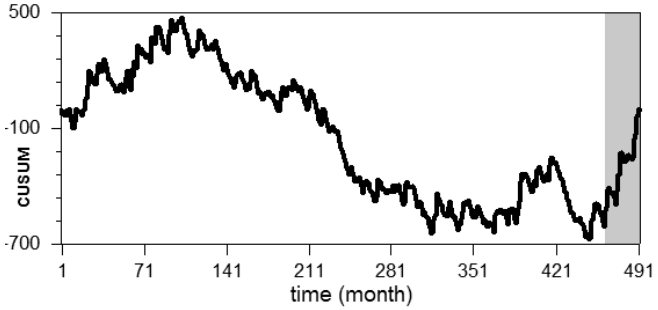

Fig.8. CUSUM of S_2

4 Models

In this section we shall present some models obtained using GEP and Box-Jenkins methods and a combination of these methods. The combination of the two is performed in a similar fashion to the technique of linear scaling employed in [13].

4.1. Models for S_1

We saw that S_1 is Gaussian, independent uncorrelated and homoscedastic, thus it is a Gaussian noise. Therefore, it is not the case to look for a better model of ARIMA type.

Using AdaGEP the best solution over 50 independent runs of each window size $w = \overline{1,5}$, had the prediction error of 64.17, and the ratio of the prediction error over the standard deviation of 0.69.

Combining the AdaGEP solution with an AR model, the result was improved. For example, starting with a model for which the prediction error was 111.874, the final model (Model 1) had the prediction error 45.94 (Fig.9), obviously an improvement over the AdaGEP solution.

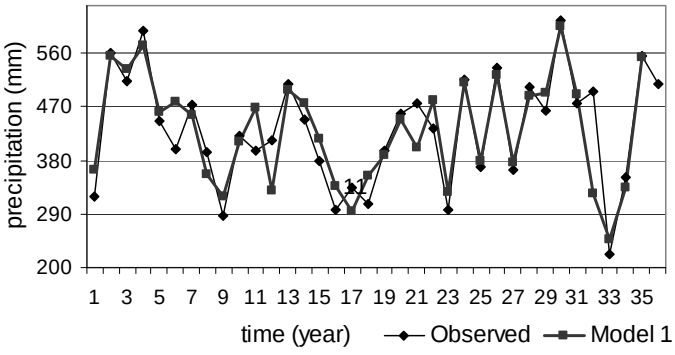

Fig.9. Model 1 for the annual data

4.2. Models for S_2

Since S_2 was normally distributed after a Box –Cox transformation, the first attempt was to use GEP and

AdaGEP to model the transformed series, but the results were not satisfactory. (Fig.10)

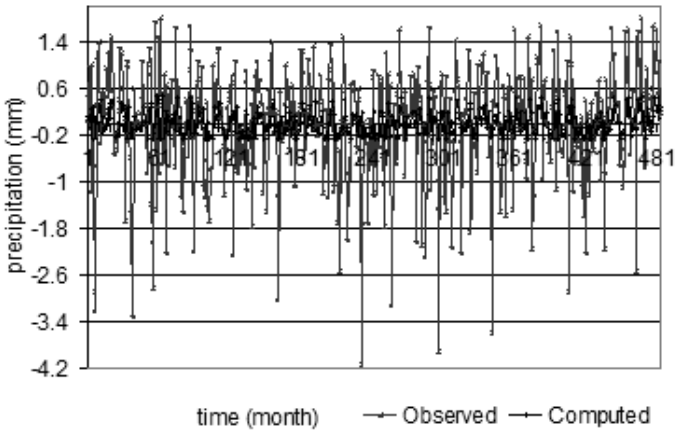

Fig.10. AdaGEP model for S_2

Therefore, for the same series, after the mean extraction a model of ARMA(2, 2) type was determined. It has the equation:

$$Z_t = 0.9577Z_{t-1} - 0.9915Z_{t-2} + \varepsilon_t - 0.929\varepsilon_{t-1} + 0.9914\varepsilon_{t-2},$$

where (ε_t) is a white noise with the variance 0.9517.

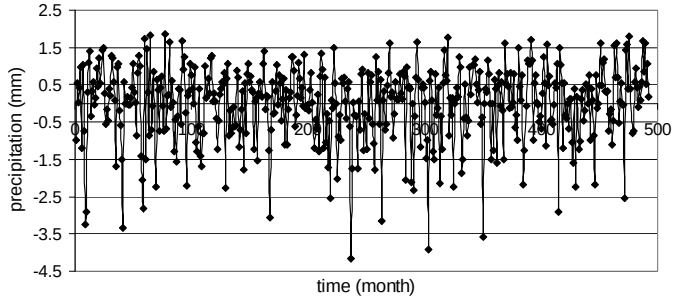

Fig.11. ARMA(2,2) model for S_2

4.2.1 **Models for S_21**

In order to model S_21, ACF and the Partial ACF were studied. Some of their values lie outside the confidence limits at the level of confidence of 95%.

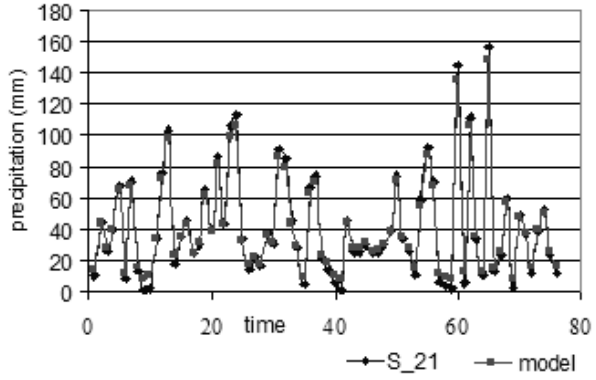

Fig.12. MA(4) model for S_21

To attenuate the differences between the data of S_21 we made a Box-Cox transformation, with λ=0.42. The form of ACF of the transformed data is of damped sine. The values of PACF are inside the confidence level, excepting the fourth. So, the model chosen was of moving average type.

Using Akaike's criterion for the model selection, the best one was:

$$X_t = \varepsilon_t - 0.2242\varepsilon_{t-4}, t \in \overline{5,76},$$

with $(\varepsilon_t)_{t\in\overline{1,76}}$ a white noise.

The charts of the best models obtained using GEP and a combination of GEP and AR are presented in Figs. 13 and 14. The corresponding errors were respectively 27 and 26.238.

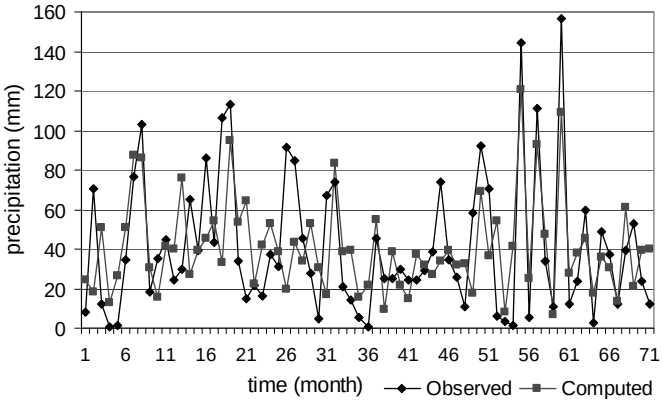

Fig.13. GEP Model for S_21

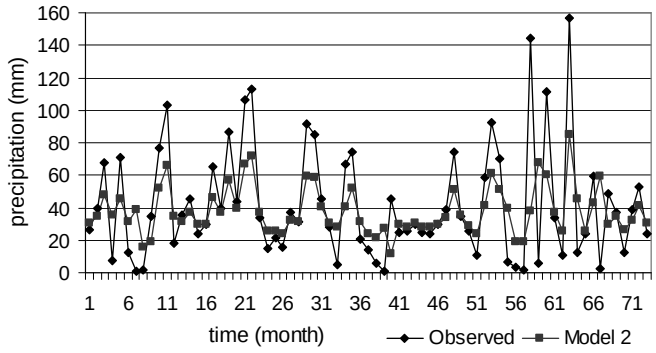

Fig.14. Combined model for S_21

4.2.2 **Models for S_22**

In Fig. 15 we present the charts of the best models obtained using genetic algorithms. We mention that the best models had the calculation error of 26.13.

A good model of ARIMA type wasn't found for this subseries. It was obtained by a decomposition process, and it is beyond the purpose of this paper.

4.2.3 **Models for S_23**

The calculated errors of the best models determined using GEP and AdaGEP were comparable with those obtained for the previous subseries.

The model determined using Box-Jenkins methods was an MA(11). The same improvement of the calculation error was registered using the combination of GEP and AR. (Fig.16)

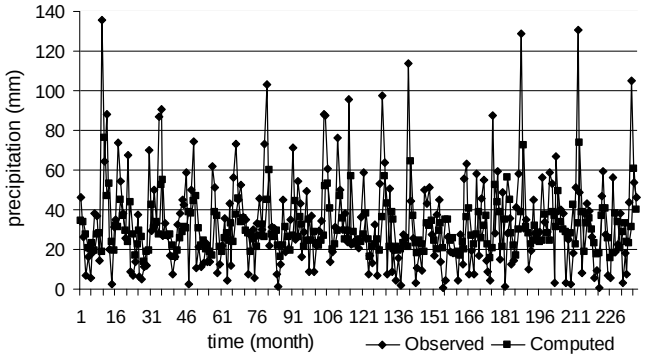

Fig.15. AdaGEP model for S_22

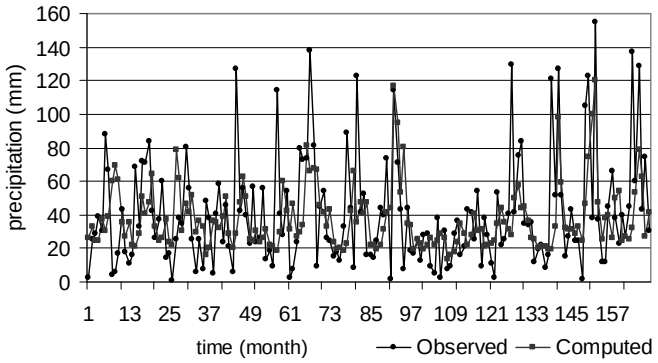

Fig.16. Combined model for S_23

4 Conclusions

The results show the adaptive gene expression programming algorithm as a fair competitor of classical methods. Better results were obtained on time series of smaller size. A straightforward explanation is the continuously changing characteristics around data that concerns weather in general, which coincides with our intuition that there exist points in meteo-hydrological time series when the underlying process changes. Our results come to support the idea that combining statistical tests for detecting change points with both heuristic methods, such as GEP, and classical Box-Jenkins methods leads to overall better models. We also reported good results combining AR with GEP. This is an idea that will be investigated in future research.

References:

[1] H. Aksoy, A, Gedikli, N. Erdem Unal, Athanasios Kehagias, Fast segmentation algorithms for long hydrometeorological time series*, Hydrological Processes*, Vol. 22, Issue 23, 2008, pp. 4600 - 4608

[2] A. Bărbulescu, *Time series with applications*, Junimea, Iasi, 2002 (in Romanian)

[3] A. Bărbulescu, E. Băutu, *Meteorological Time Series Modeling Based on Gene Expression Programming*, submitted

[4] E. Băutu, A. Băutu, H. Luchian, AdaGEP - An. Adaptive Gene Expression Programming, *Proceedings of the Ninth international Symposium on Symbolic and Numeric Algorithms For Scientific Computing (September* 26-29, 2007), SYNASC, IEEE Computer Society, 2007, pp. 403-406.

[5] P. Brockwell, R. Davies, *Introduction to time series*, Springer, New York, 2002

[6] T. A. Buishard, Tests for detecting a shift in the mean of hydrological time series, *Journal of Hydrology*, Vol.73, 1984, pp. 51-69

[7] A. Busuioc, H. von Storch, Conditional stochastic model for generating daily precipitation time series, *Climate Research*, Vol. 24, 2003, pp. 181–195

[8] S. P. Charles, B. C. Bates, I. N. Smith, J.P. Hughes, Statistical Downscaling of observed and modeled atmospheric fields, *Hydrological Processes*, Vol. 18, No. 8, 2004, pp. 1373-1394,

[9] R. A. Davis, T. C. M. Lee, G. A. Rodriguez-Yam, Structural breaks estimation for non-stationary time series signals, *Journal of the American Statistical Association*, Vol. 101, No., 473, 2006, pp. 223–229

[10] C. Ferreira, *Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence*, Springer - Verlag, 2006

[11] P. Hubert, The segmentation procedure as a tool for discrete modeling of hydrometeorological regimes, *Stochastic Environmental Research and Risk Assessment*, Vol.14, pp 297-304, 2000

[12] N. Q. Viet Hung, D. T. Anh, Combining SAX and Piecewise Linear Approximation to Improve Similarity Search on Financial Time Serie, *IEEE International Symposium on Information Technology Convergence*, 2007. ISITC 2007, pp. 58 – 62

[13] Keijzer, M. 2004. Scaled Symbolic Regression. *Genetic Programming and Evolvable Machines* 5, 3 (Sep. 2004), pp. 259-269.

[14] A. N. Pettitt, A non-parametric approach to the change-point problem, *Applied Statistics*, Vol. 28, No. 2, 1979, pp. 126 - 135.

[15] D.J. Seskin, *Handbook of parametric and nonparametric statistical procedures*, Chapmann & Hall/CRC, Boca Raton, 2007

[16] N. Wagner, Z. Michalewicz, M. Khouja, and R. R. Mcgregor, Time series forecasting for dynamic environments: The Dyfor genetic program model. *IEEE Transactions on Evolutionary Computation*, 11(4):433–452, 2007

[18] Yuelong Zhu, De Wu and Shijin Li, A Piecewise Linear Representation Method of Time Series Based on Feature Points, *Lecture Notes In Computer Science,* Vol. 4693, 2007, pp. 1066-1072

Acknowledgements: This article was supported by grant PNII ID_262 and the project TOMIS (CNMP PNCDI-2 11-041/2007).