Short-term Rainfall Time Series Prediction with incomplete data

Cristian Rodriguez Rivero
LIMAC - Department of Electronic Engineering
Universidad Nacional de Córdoba
Córdoba, Argentina
crodriguezrivero@efn.uncor.edu

Hector Daniel Patiño
INAUT - Advanced Intelligent Systems Laboratory
Universidad Nacional de San Juan
San Juan, Argentina
dpatino@inaut.unsj.edu.ar

Julian Antonio Pucheta
LIMAC - Department of Electronic Engineering
Universidad Nacional de Córdoba
Córdoba, Argentina
jpucheta@efn.uncor.edu

***Abstract*—In order to predict short-term times series with incomplete data, a proposed approach is presented based on the energy associated of series. A benchmark of rainfall time series and Mackay Glass (MG) samples are used. An average smoothing technique is adopted to complete the dataset. The structure of the predictor filter is changed taking into account the energy associated of the short series. The H parameter is used to estimate the roughness of the complete series, the real and forecasted one. The next 15 values are used as validation and horizon of the time series presented by series of cumulative monthly historical rainfall from La Sevillana, Cordoba, Argentina and samples of the Mackay Glass (MG) differential equation. The performance of the proposed filter shows that even the short dataset is incomplete, besides a linear smoothing technique employed, the prediction is almost fair. Although the major result shows that the predictor system based on energy associated to series has an optimal performance from several samples of MG equations and, in particular, MG1.6 and SEV rainfall time series, this method provides a good estimation when the short-term series are taken from one point observations.**

Keywords— Incomplete data; energy associated to series; neural networks; time series prediction; nonlinear systems.

I. INTRODUCTION

Time series forecasting recently has a preponderant significance in order to know which will be the best the behavioral of a system in study such as the availability of estimated scenarios for water predictability [8], the rainfall forecast problem [19] [23] [24], [20] in some geographical points of Cordoba, the energy demand purposes [4], the guidance of seedling growth [13]. For general feed-forward neural networks [16] [17] [18], the computational complexity of these solutions grows exponentially with the number of missing features.

The motivation of this work arises out of incomplete data that poses a difficulty to the analysis and decision making processes which depend on this data, requiring methods of estimation which are accurate and efficient. This work describes a linear method for approximation problem of missing information, which is applicable to a large class of learning algorithms [2], [3], including ANNs. One major advantage of the proposed solution is that the complexity does not increase with an increasing number of missing inputs. The solutions can easily be generalized to the problem of uncertain (noisy) inputs.

Various techniques exist as a solution to this problem, ranging from data deletion to methods employing statistical and artificial intelligence techniques to impute for missing variables. However, a linear estimation is employed. Furthermore, we make assumptions about the data that may not be true, affecting the quality of decisions made based on this data. The estimation of incomplete data in vector elements in real time processing applications requires a system that possesses the knowledge of certain characteristics such as correlations between variables, which are inherent in the input space [12]. Those are taken from the Mackay Glass benchmark equation and cumulative historical rainfall whose forecast is simulated by a Monte Carlo approach employing ANN.

The main contribution here is the forecast system based on energy associated to series that uses incomplete data for tuning its parameters at the same time the historical recorded data is relatively short. The filter parameter is put in function of the roughness of the short time series, between its smoothness. In addition, this forecasting tool is intended to be used by farmers to maximize their profits, avowing profit losses over the misjudgment of future movements to maximize their utilities. A one-layered feed-forward neural network, trained by the Levenberg-Marquardt algorithm is implemented in order to give the next 15 values.

II. METHODOLOGY AND THEORY

A. Overview on fractional Brownian motion

In this work the Hurst's parameter is used in the learning process to modify on-line the number of patterns, the number of iterations, and the number of filter's inputs. This *H* serves to have an idea of roughness of a signal, and to determine its stochastic dependence. The definition of the Hurst's parameter appears in the Brownian motion from generalizing the integral

to a fractional one. The Fractional Brownian Motion (*fBm*) is defined in the pioneering work by Mandelbrot [22] through its stochastic representation:

$$B_H(t)=\frac{1}{\Gamma\left(H+\frac{1}{2}\right)}\left(\begin{array}{l}\int_0^{\infty}\left((t-s)^{H-\frac{1}{2}}-(-s)^{H-\frac{1}{2}}\right)\\ +\int_0^{\infty}(t-s)^{H-\frac{1}{2}}dB(s)\end{array}\right)dB(s) \quad (1)$$

where, $\Gamma(\cdot)$ represents the Gamma function

$$\Gamma(\alpha)=\int_0^{\infty}x^{\alpha-1}e^{-x}dx \quad (2)$$

and $0<H<1$ is called the Hurst parameter. The integrator B is a stochastic process, ordinary Brownian motion. Note, that B is recovered by taking $H=1/2$ in (1). Here, it is assumed that B is defined on some probability space (Ω, F, P), whereΩ, F and P are the sample space, the sigma algebra (event space) and the probability measure respectively. Thus, an *fBm* is a time continuous Gaussian process depending on the so-called Hurst parameter $0<H<1$. The ordinary Brownian motion is generalized to $H=0.5$, and whose derivative is the white noise.

The *fBm* is self-similar in distribution and the variance of the increments is defined by:

$$Var\left(B_H(t)-B_H(s)\right)=\nu\left|t-s\right|^{2H} \quad (3)$$

where ν is a positive constant.

B. Data treatment

The proposed work shows the next fifteen short time series prediction values in the field of meteorological variables such as cumulative rainfall [14]. The dataset chosen is from historical data 2004 to 2011 from La Sevillana establishment, located at Cordoba, Argentina shown in Fig. 1.

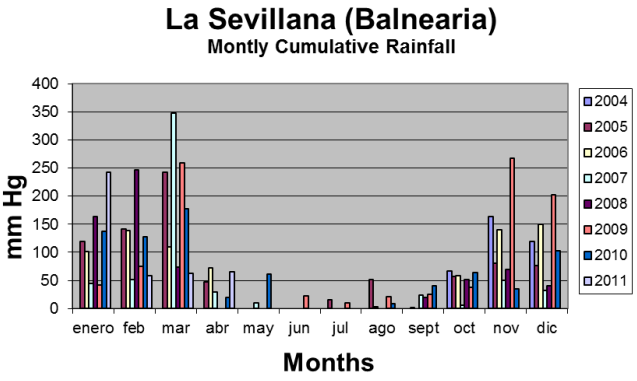

Fig. 1 Cumulative Monthly Rainfall of La Sevillana (SEV)

The original dataset (SEV) used is incomplete and contains 79 data of cumulative monthly rainfall data, in which there are 14 months values incomplete resulting in a non-determinist series, respectively. This kind of behavior is difficult to predict because seasonality is not well-determined by few data. For the sake of making a fair prediction, a linear smoothing was employed to replace the incomplete data. This consists of averaging on vertical column shows in Fig. 2, the prior and posterior value that corresponds to the same year.

C. Mackay-Glass time series

The second benchmark of series is obtained from solution of the MG equation. This equation serves to model natural phenomena and has been used in earlier work to implement different methods of comparison to make forecast [6], which is explained by the time delay differential MG equation [22], defined as

$$\dot{y}(t)=\frac{\alpha y(t-\tau)}{1+y^{c}(t-\tau)}-\beta y(t) \quad (4)$$

where α, β varies and c=10 are parameters and τ=100 is the delay time. According as τ increases, the solution turns from periodic to chaotic. Thereby, a time series with a random-like behavior is obtained, and the long-term behavior changes thoroughly by changing the initial conditions to obtain the stochastic dependence of the deterministic time series according to its roughness [6].

In this work the Hurst's parameter is used in the learning process to modify on-line the number of patterns, the number of iterations, and the number of filter's inputs of the ANN. This H serves to have an idea of roughness of a signal [21] and the time series are considered as a trace of an fBm depending on the so-called Hurst parameter $0<H<1$. The MG benchmark chosen are called MG085, MG1.6 and MG1.9.

	ja	feb	mar	ap	may	ju	jul	au	se	oc	nov	dic	Annual
2004	x	x	x	x	x	x	x	x	x	67	163	119	**349**
2005	119	142	242	47	0	0	15	52	2	57	80	77	**833**
2006	101	139	110	72	0	0	0	3	0	58	140	150	**773**
2007	45	52	348	30	10	0	0	0	24	6	50	32	**597**
2008	163	247	74	x	x	x	x	x	20	52	70	40	**666**
2009	42	75	259	0	0	22	10	21	25	37	267	202	**960**
2010	137	128	177	20	61	0	0	8	40	64	35	103	**773**
2011	243	59	63	65	x	x	x	x	x	x	x	x	**430**
Average	**121**	**120**	**182**	**39**	**14**	**4**	**5**	**17**	**19**	**49**	**115**	**103**	**788**

Fig. 2 Average technique adopted to complete the rainfall dataset.

III. Problem Formulation

The main issue when forecasting a time series is how to retrieve the maximum of information from the available data. In this case, the lack of data in the dataset is taken into account in order to predict one step ahead for the filter based on ANN. It is proposed to fill the absence of data by using prior and posterior data. Four dataset are built following Fig. 2. In the first one, the lack data is completed by taking the same ensemble of data of the past year. The second one by using the same ensemble of the next year, the third one is completed with zeros and lastly is filled in by averaging the prior and posterior year. The same analogy is used to construct MG05 and MG1.8 dataset solution of (1).

The coefficients of the ANNs filter are adjusted on-line in the learning process, by considering a criterion that modifies

at each pass of the time series the number of patterns, the number of iterations and the length of the tapped-delay line, in function of the Hurst's value (*H*) calculated from the time series according to the stochastic behavior of the series, respectively.

In this work, the present value of the time series is used as the desired response for the adaptive filter and the past values of the signal serve as input of the adaptive filter [11]. Then, the adaptive filter output will be the one-step prediction signal. In the block diagram of the nonlinear prediction scheme based on an ANN filter is shown. Here, a prediction device is designed such that starting from a given sequence $\{x_n\}$ at time n corresponding to a time series it can be obtained the best prediction $\{x_e\}$ for the following sequence of 15 values. Hence, it is proposed a predictor filter with an input vector l_x, which is obtained by applying the delay operator, Z^{-1}, to the sequence $\{x_n\}$. Then, the filter output will generate x_e as the next value, that will be equal to the present value x_n. So, the prediction error at time k can be evaluated as

$$e(k) = x_n(k) - x_e(k) \tag{5}$$

which is used for the learning rule to adjust the ANN weights.

IV. Proposed Approach to Calculate the Energy Associated of Series

A. Approximation by Primitive of Integration

The area resulting of integrating the data time series of MG and rainfall data series is the primitive, that is obtained by considering each value of time series its derivate [5];

$$\int_{t_k}^{t_{k+1}} y_t dt \cong y_t \left(t_{k+1} - t_k\right) \tag{6}$$

where y_t is the original value time series. The area approximation by its periodical primitive is:

$$I_{t_n} = \int_{t_n}^{t_{n+p}} y_t dt = Y_t\Big|_{t_n}^{t_{n+p}}, n = 1,2,...N. \tag{7}$$

During the learning process, those primitives are calculated as a new entrance to the ANN, in which the prediction attempts to even the area of the forecasted area to the primitive real area predicted. The real primitive integral is used in two instances, firstly from the real time series an area is obtained and run by the algorithm proposed. The *H* parameter from this time series is called H_A. On the other hand, the data time series is also forecasted by the algorithm, so the *H* parameter from this time series is called H_S. Finally, after each pass the number of inputs of the nonlinear filter is tuned —that is the length of tapped-delay line, according to the following heuristic criterion. After the training process is completed, both sequences - $\{\{I_n\},\{I_e\}\}$ and $\{\{y_n\},\{y_e\}\}$ - in accordance with the hypothesis that should have the same *H* parameter. If the error between H_A and H_S is greater than a threshold parameter θ the value of l_x is increased (or decreased), according to $l_x \pm 1$. Explicitly,

$$l_x = l_x + sign(x) \tag{8}$$

Here, the threshold θ was set about 1%.

V. Prediction Results

A. Generations of areas from benchmark

Primitives of time series are obtained from sampling the MG equations with parameters shown in Table I, with τ=100, c=10 and varying β, α. This collection of coefficients was chosen to generate time series whose *H* parameters vary between 0 and 1 [15] and [16]. In fact, the chosen one was selected in accordance with its high roughness.

TABLE I. Parameters to generate the MG times series

Series No.	β	α	c	H
MG0.85	0.85	20	10	0.21
MG1.6	1.6	20	10	0.029
MG1.9	1.6	30	10	0.24

TABLE II. Parameter of La Sevillana Rainfall series

Series No	H
SEV	0.28

B. Performance measure for forecasting

In order to test the proposed design procedure of the ANN - based nonlinear predictor, an experiment with time series obtained from the MG solution was performed. The performance of the filter is evaluated using the Symmetric Mean Absolute Percent Error (*SMAPE*) proposed in the most of metric evaluation, defined by

$$SMAPE_S = \frac{1}{n}\sum_{t=1}^{n} \frac{\left|X_t - F_t\right|}{\left(X_t + F_t\right)/2} \cdot 100 \tag{9}$$

where *t* is the observation time, n is the size of the test set, **s** is each time series, X_t and F_t are the actual and the forecasted time series values at time t respectively. The *SMAPE* of each series s calculates the symmetric absolute error in percent between the actual X_t and its corresponding forecast value F_t, across all observations *t* of the test set of size n for each time series **s**.

C. Forecasting Results

Each time series is composed by samples of MG solutions and La Sevillana rainfall time series. Three classes of data sets are used. The first one is the original time series used by the algorithm to train the predictor filter, which comprises 64 values. The next one is the primitive obtained by integrating the original time series data. The last one is used to compare if the forecast is acceptable or not, in which the last 15 of 79 values can be used to validate the performance of the prediction system. A comparison of roughness is made between the rainfall time and MG series.

The Monte Carlo method was used to forecast the next 15 values from La Sevillana rainfall series (SEV), MG085, MG1.6 and MG1.9 time series and their primitive. Such outcomes are shown from Fig. 2 to Fig. 5. The plot shown in Fig. 2 and 4 are from H dependent ANN. Fig 3 and Fig 5 are obtained by the proposed approach.

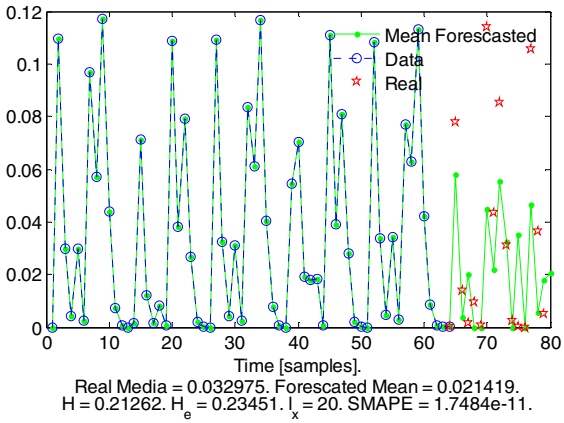

Fig. 3. ANN algorithm for MG085.

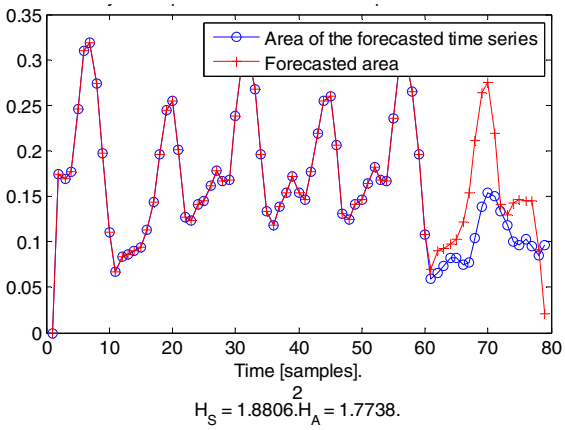

Fig. 4. Primitive of MG085 time series.

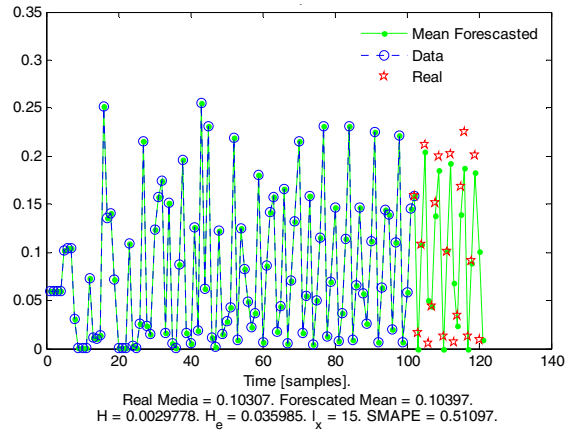

Fig. 5. ANN algorithm for MG1.6.

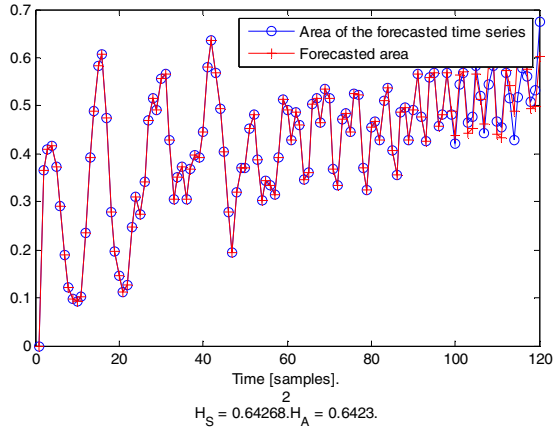

Fig. 6. Primitive of MG1.6 time series.

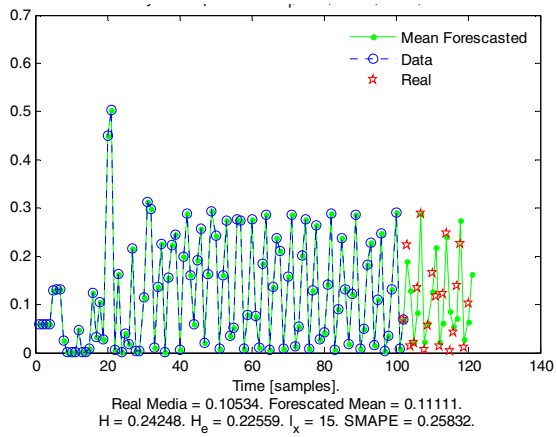

Fig. 7. ANN H independent algorithm for MG1.9.

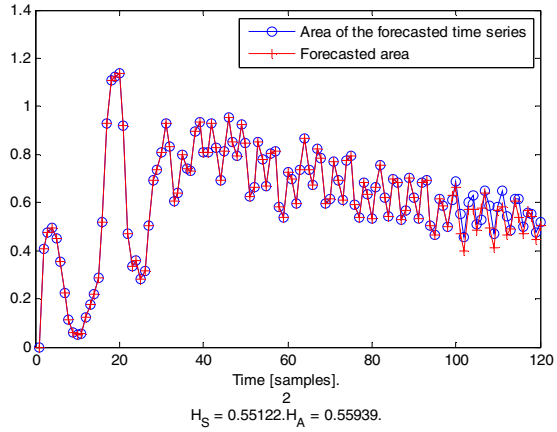

Fig. 8. Primitive of MG1.9 time series.

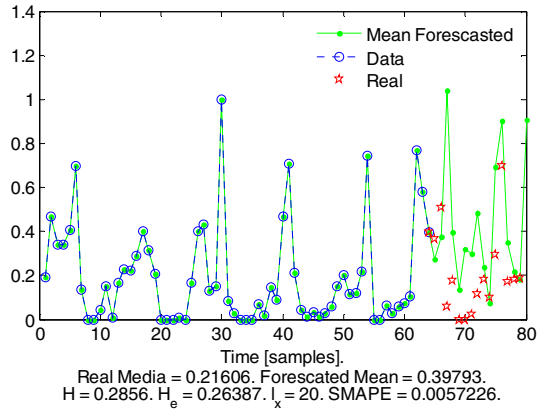

Fig. 9. ANN algorithm for La Sevillana rainfall series.

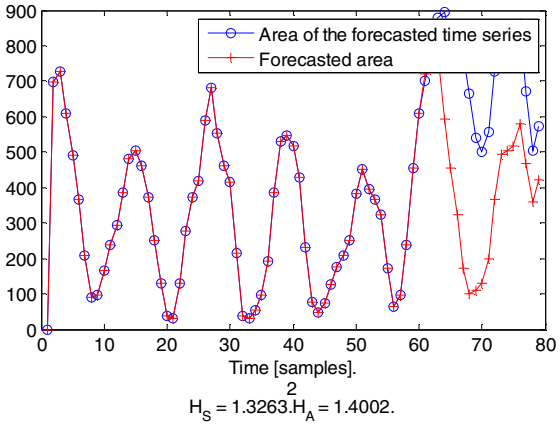

Fig. 1. Fig. 10. Primitive of La Sevillana rainfall series.

TABLE III. TABLE III. COMPARISONS OBTAINED BY THE PROPOSED APPROACH

Series No.	H_S	H_A	SMAPE
Fig.3	1.8806	1.7738	1.74e-11
Fig.5	0.6426	0.6423	0.510
Fig.7	0.5512	0.5593	0.258
Fig. 9	1.3226	1.4002	5e-3

The algorithm achieves the long or short term stochastic dependence measured by the Hurst parameter in order to make more precisely the prediction. The forecasted time series area is put as a new entrance to the ANN and serves to be compared with the real primitive obtained of the time series.

The figures show a class of high roughness time series selected from a benchmark of MG Equation and compared with SEV rainfall series. These are classified by their statistically dependency, so the algorithm is adjusted by depending on the *H* parameter. At Table III shows a good performance when it takes into accounts the roughness of the series considering the use of the stochastic dependence measured by the *H* parameter.

CONCLUSIONS

In this work, short-term rainfall time series prediction with incomplete data by means of energy associated of series was presented. The learning rule proposed to adjust the ANNs weights is based on the Levenberg-Marquardt method and energy associated to series as a new input. Likewise, in function of the short-term stochastic dependence of the time series evaluated by the Hurst parameter *H*, the performance of the proposed filter shows that even the short dataset is incomplete, besides a linear smoothing technique employed, the prediction is almost fair. The major result shows that the predictor system based on energy associated to series has an optimal performance from several samples of MG equations and, in particular, MG1.6 and SEV rainfall time series. These were considered as a path of a fractional Brownian motion [7] whose *H* parameter measured belongs to a class of high roughness signal, which is assessed by H_S and H_A, respectively. This approach encourages forecasting meteorological variables such as moisture soil series, daily and hour rainfall and water runoff when the observations are taken from a single point.

ACKNOWLEDGMENT

This work was supported by Universidad Nacional de Córdoba (UNC), FONCYT-PDFT PRH N°3 (UNC Program RRHH03), SECYT UNC, Universidad Nacional de San Juan – Institute of Automatics (INAUT), National Agency for Scientific and Technological Promotion (ANPCyT) and Departments of Electrotechnics – Electrical and Electronic Engineering - Universidad Nacional of Cordoba.

REFERENCES

[1] Abry, P.; P. Flandrin, M.S. Taqqu, D. Veitch. (2003), Self-similarity and long-range dependence through the wavelet lens. Theory and applications of long-range dependence, Birkhäuser, pp. 527-556.

[2] Bishop, C. Pattern Recognition and Machine Learning. Boston: Springer, 2006.

[3] Bishop, C. (1995). Neural Networks for Pattern Recognition. University Press. Oxford

[4] E.Toth, A.Brath, A.Montanari,"Comparison of short-term rainfall prediction models for real-time flood forecasting", Journal of Hydrology 239 (2000) 132–147.

[5] C. Rodríguez Rivero, M. Herrera, J. Pucheta, J. Baumgartner, D. Patiño and V. Sauchelli "High Roughness Time Series Forecasting based on energy associated of series", Journal of Communication and Computer, Vol. 9 No. 5 2012, Pp 576-586, ISSN 1548-7709, USA, David Publishing Company.

[6] Nakama, T. (2009). Theoretical analysis of batch and on-line training for gradient descent learning in neural networks. Neurocomputing, 73, 151-159.

[7] Flandrin, P., (1992) Wavelet analysis and synthesis of fractional Brownian motion" IEEE Trans. on Information Theory, 38, pp. 910-917.

[8] N. Q. Hung, M. S. Babel, S. Weesakul, and N. K. Tripathi "An Artificial Neural network Model for rainfall Forercasting in Bangkok,Thailand", Hydrol. Earth Syst. Sci., 13, 1413–1425, 2009.

[9] Mandelbrot, B. B., (1983), The Fractal Geometry of Nature, Freeman, San Francisco, CA. 1983.

[10] Masulli, F., Baratta, D., Cicione, G., Studer, L. "Daily Rainfall Forecasting using an Ensemble Technique based on Singular Spectrum Analysis". In Proceedings of the International Joint Conference on Neural Networks IJCNN 01, pp. 263-268, vol. 1, IEEE, Piscataway, NJ, USA, 2001.

[11] A. Khosravi, S. Nahavandi, D. Creighton, and A. F. Atiya, "Comprehensive review of neural network-based prediction intervals and new advances," *IEEE Trans. Neural Netw.*, vol. 22, no. 9, pp. 1341–1356, Sep. 2011.

[12] Pucheta, J., Patiño, H. D., Kuchen, B. (2008) "A Statistically Dependent Approach for the Monthly Rainfall Forecast from One Point Observations". In Proc. of the Second IFIP Conference on Computer

and Computing Technologies in Agriculture (CCTA2008) October 18-20. Beijing, China. 2008.

[13] Pucheta, J., Patiño, H., Schugurensky, C., Fullana, R., Kuchen, B. Optimal Control Based-Neurocontroller to Guide the Crop Growth under Perturbations. Dynamics Of Continuous, Discrete And Impulsive Systems Special Volume Advances in Neural Networks-Theory and Applications. DCDIS A Supplement, Advances in Neural Networks, Watam Press, Vol. 14(S1), pp. 618—623. 2007.

[14] Julián A. Pucheta , Cristian M. Rodríguez Rivero , Martín R. Herrera, Carlos A. Salas, H. Daniel Patiño y Benjamín R. Kuchen, "A NN approach for cumulative monthly rainfall time series forecasting tuned by roughness", International Journal of the Physical Sciences, Academic Journals, ISSN 1992 – 1950, International Journal of the Physical Sciences, Academic Journals..

[15] Velásquez Henao, Juan David, Dyna, Red. Pronóstico de la serie de Mackey glass usando modelos de regresión no-lineal. Universidad Autónoma de Mexico. Campus Aragón. 2004.

[16] Zhang, G.; B.E. Patuwo, and M. Y. Hu. "Forecasting with artificial neural networks: The state of art". J. Int. Forecasting, vol. 14, pp. 35-62. 1998.

[17] J. Pucheta, M., C. Rodríguez Rivero, M. Herrera, C. Salas, D. Patiño and B. Kuchen. "A Feed-forward Neural Networks-Based Nonlinear Autoregressive Model for Forecasting Time Series". Revista Computación y Sistemas, Centro de Investigación en Computación-IPN, México D.F., México, Computación y Sistemas Vol. 14 No. 4, pp. 423-435 ISSN 1405-5546, 2011.

[18] C. Rivero Rodríguez, J. Pucheta, J. Baumgartner, H.D. Patiño and B. Kuchen, "An Approach for Time Series Forecasting by simulating Stochastic Processes Through Time-Lagged feed-forward neural network". The 2010 World Congress in Computer Science, Computer Engineering, and Applied computing. Las Vegas, Nevada, USA, July 12-15, 2010. DMIN'10 Proceedings ISBN 1-60132-138-4 CSREA Press,p.p 278, (CD ISBN 1-60132-131-7), USA, (2010).

[19] C. Rodríguez Rivero, M. Herrera, J. Pucheta, J. Baumgartner, D. Patiño and V. Sauchelli "High Roughness Time Series Forecasting based on energy associated of series", Journal of Communication and Computer, Vol. 9 No. 5 2012, pp 576-586, ISSN 1548-7709, USA, David Publishing Company.

[20] C. Rodriguez Rivero, J. Pucheta, H. Patiño, J. Baumgartner, S. Laboret and V. Sauchelli. "Analysis of a Gaussian Process and Feed-Forward Neural Networks based Filter for Forecasting Short Rainfall Time Series" . 2013 International Joint Conference on Neural Networks, Texas, USA. Print Edition: IEEE Catalog Number: CENSUS-ART, ISBN: 978-1-4673-6129-3, ISSN: 2161-4407, CD Edition: IEEE Catalog Number: CFPlSUS-CDR, ISBN: 978-1-4673-6128-6. 2013.

[21] Dieker, T. (2004). Simulation of fractional Brownian motion. The Netherlands MSc theses, University of Twente Amsterdam.

[22] Glass L., Mackey M. C. (1998). From Clocks to Chaos, The Rhythms of Life. Princeton, NJ: Princeton University Press.

[23] C. Rodríguez Rivero, M. Herrera, J. Pucheta, J. Baumgartner, D. Patiño and V. Sauchelli and S. Laboret, "Time series forecasting using Bayesian method: application to cumulative rainfall"ISSN 1548-0992. Pp. 359 364. IEEE LATIN AMERICA TRANSACTIONS, VOL. 11, NO. 1, FEB. 2013.

[24] Julian Pucheta, Cristian Rodriguez Rivero, Martín Herrera, Carlos Salas, Victor Sauchelli, "Rainfall forecasting using sub sampling non-parametric methods", ISSN 1548-0992. Pp. 346-350. IEEE LATIN AMERICA TRANSACTIONS, VOL. 11, NO. 1, FEB. 2013.