

Daily Reservoir Inflow Forecasting Using Fuzzy Inference Systems

Ivette Raymunda Luna Huamaní and
Rosangela Ballini
Institute of Economics
UNICAMP
Campinas, Brasil
{ivette.ballini}@eco.unicamp.br

Ieda Geriberto Hidalgo
Faculty of Technology
UNICAMP
Limeira, Brasil
iedahidalgo@ft.unicamp.br

Paulo Sérgio Franco Barbosa and
Alberto Luiz Francato
Faculty of Civil Engineering
UNICAMP
Campinas, Brasil
{franco, francato}@fec.unicamp.br

Abstract— This paper presents the application of a methodology for daily reservoir inflow forecasting in Brazilian hydroelectric plants. The methodology is based on Fuzzy Inference Systems (FIS) and the technique used for adjusting of the model parameters is an offline version of the Expectation Maximization (EM) algorithm. In order to automate the application of the methodology and facilitate the analysis of the results, a tool that allows managing streamflow forecasting studies and visualizing their information in graphical form was developed. A case study was applied to the data from three Brazilian hydroelectric plants whose operation is under the coordination of the Electric System National Operator. They are located in the Grande basin, a part of the Parana basin with two main rivers: the Grande and the Pardo. The benefits of the model are analyzed using statistics calculations, such as: root mean square error, mean absolute percentage error, mean absolute error and mass curve coefficient. Besides that, graphics that compare the registered and predicted streamflow are presented. The results show an adequate performance of the model, leading to a promising alternative for daily streamflow forecasting.

Keywords: *Inflow forecasting; Fuzzy systems; Hydroelectric plants; Reservoir management; Computacional tool.*

I. INTRODUCTION

One of the great economic and environmental challenges facing humanity nowadays is the production of electricity. In Brazil, the power generation is assigned to the task of planning and scheduling of hydroelectric plants. The objective of this task is to ensure an economic and reliable operation policy for the system.

Forecasting of streamflows is an important step for planning and scheduling of hydroelectric power system. However, this step is still one of the most difficult issues in hydrological science because it is characterized as a problem dynamic, uncertain and nonlinear. The purpose of forecasting is to support decision-making, help operate reservoirs, assess resource capabilities and determine pricing for the out-of-province sale of energy.

Several researches have been devoted to the formulation and development of models that aims to improve the quality of hydrological prediction. In general, the models are assigned to

one of three broad categories: *deterministic, conceptual* or *parametric*.

Deterministic models describe the rainfall-runoff process using physical laws of mass and energy transfer. Conceptual models provide simplified representations of key hydrological process using a perceived system. Parametric models use mathematical transfer functions to relate meteorological variables to runoff [1].

Deterministic and conceptual models require a large amount of high-quality data associated with hydrological, meteorological and natural geographical characteristics as well as human activities. Parametric models do not presuppose a detailed understanding of the basin's physical characteristics, nor does it require extensive data preprocessing [2]. Therefore, neural network and fuzzy systems are particularly powerful in situation where it is difficult to determine the physical process or when it is not possible to obtain a physical interpretation of the mathematical model representation [3]. Their potential for inflow forecasting is demonstrated in [4], [5] and [6].

In [7], an offline Fuzzy Inference System (FIS) was used for predicting annual inflows that are disaggregated in the respective monthly samples used for long-term hydropower scheduling. The use of these tools for the daily or hourly forecasting may be an interesting alternative that complements time series analysis for inflow modeling, usually performed in hydroinformatics via physical and conceptual models [8].

Given the importance of streamflow forecasting for power generation systems, this paper presents a Takagi-Sugeno (TS) FIS for daily inflow forecasting. The learning algorithm is based on the offline version of the Expectation Maximization (EM) algorithm. The objective is to improve the quality of this information, contributing to the choice of an operational policy that meets the consumer market in an economic and safe way.

The context of this paper is organized as follow. Initially, Section II presents the model general structure and the optimization algorithm. Then, in Section III a case study is shown. This section is divided in four items. The first describes the study area and the data collection; in the second item the methodology and the modeling are explained; the third shows the tool used in this work and the fourth item presents the

results and discussions. After that, in Section IV the conclusions are presented.

II. MODEL DESCRIPTION

Model detailed is already used in different applications involving time series forecasting. It was originally proposed as a model initialization of the adaptive fuzzy system model proposed in [8]. In order to increase the clarity of this paper a general description of the model structure is presented next.

A. General Structure

The model structure adopted, as its name indicates, is mainly based on the first order TS-FIS [9].

Let denote the input vector at instant time k as $x^k = [x_1^k \ x_2^k \ \dots \ x_p^k]$. The output model for the correspondent input x^k is denoted by y^k . The input space represented $x^k \in R^p$ is partitioned into M sub-regions, each represented by a fuzzy rule (Fig. 1).

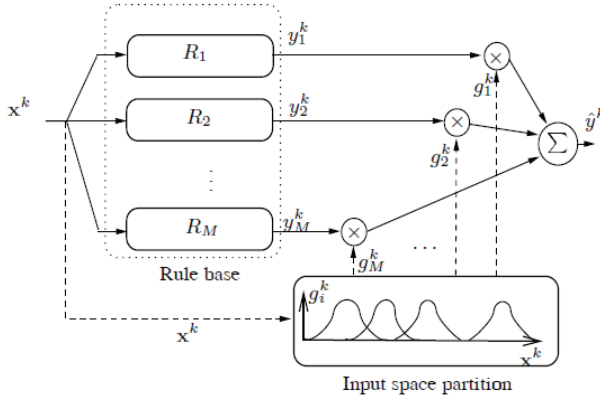


Figure 1. General model structure.

The antecedents of each If-Then fuzzy rule (R_i) are represented by their respective centers $c_i \in R^p$ and covariance matrices $V_i |_{p \times p}$. The consequents are represented by local linear models, with outputs y_i^k , $i = 1, \dots, M$ defined by:

$$y_i^k = \phi^k \times \theta_i^T \quad (1)$$

where $\phi^k = [1 \ x_1^k \ x_2^k \ \dots \ x_p^k]$ is the extended input vector where the first component is increased for considering a constant term in the local linear model; $\theta_i = [\theta_{i0} \ \theta_{i1} \ \dots \ \theta_{ip}]$ is the coefficient vector of the local linear model related to the i^{th} rule.

Each input pattern has a membership degree associated with each region of the input space partition. This is calculated through membership functions $g_i(x^k) \in [0,1]$ that vary according to centers and covariance matrices related to the fuzzy partition, and are computed by:

$$g_i(x^k) = g_i^k = \frac{\alpha_i P[i | x^k]}{\sum_{q=1}^M \alpha_q P[q | x^k]} \quad (2)$$

where α_i are positive coefficients satisfying $\sum_{i=1}^M \alpha_i = 1$ and $P[i | x^k]$ is defined according to:

$$P[i | x^k] = \frac{1}{(2\pi)^{p/2} \det(V_i)^{1/2}} \exp(-(x^k - c_i)' V_i^{-1} (x^k - c_i)) \quad (3)$$

where $\det(\cdot)$ is the determinant function. The final model output $\hat{y}^k$ is calculated by a non-linear weighted average of the local outputs and its respective membership degrees. Thus, the predicted value of the output variable for future time instant k is given by:

$$\hat{y}^k = \sum_{i=1}^M g_i^k y_i^k \quad (4)$$

B. Optimization Algorithm

The EM algorithm has been used for the optimization of several computational intelligence based methods, such as the works presented in [8], [10] and [11]. Therefore, this work suggests the use of the offline EM algorithm for the optimization of the FIS parameters.

First, the FIS structure is initialized. In order to define the number of rules M to codify in the model structure, the unsupervised Subtractive Algorithm (SC) is used over the input-output training data set [12]. Initial values for spreads codified in the covariance matrix are the same used by the SC algorithm. Therefore, the FIS structure is initialized as follows:

- $c_i^0 = \psi_i^0 |_{1 \dots p}$ where $\psi_i^0 |_{1 \dots p}$ is composed of the first p components of the i^{th} center found by the SC algorithm;
- $\sigma_i^0 = 1.0$;
- $\theta_i^0 = [\psi_i^0 |_{p+1} \ 0 \dots 0]_{1 \times p+1}$, where $\psi_i^0 |_{p+1}$ is the $(p+1)^{\text{th}}$ component of the i^{th} center found by the SC algorithm;
- $V_i^0 = r_a^2 I$, where I is a $p \times p$ identity matrix and r_a is the spread parameter used by the SC algorithm;
- $\alpha_i^0 = 1/M$.

After this initialization, model parameters are readjusted on the basis of the offline EM algorithm, (see [13] for the complete formulation), with the objective of maximizing the log-likelihood $\mathfrak{S}$ of the observed values of y^k at each M step of the learning process. This objective function is defined by:

$$\mathfrak{J}(D, \Omega) = \sum_{k=1}^N \ln \left(\sum_{i=1}^M g_i(x^k, C) \times P(y^k | x^k, \theta_i) \right) \quad (5)$$

where N is the number of input-output patterns used for model optimization, $D = \{x^k, y^k \mid k = 1, \dots, N\}$, Ω contains all model parameters and C contains just the antecedents parameters (centers and covariance matrices). As noticed, for maximizing $\mathfrak{J}$, it is necessary to know the data distribution. Due to the lack of knowledge of the real distribution of precipitation and inflow data, the model is adjusted assuming a normal distribution of the observed records.

III. CASE STUDY

A. Study Area and Data Collection

The case study considers a part of the Paraná basin, located in the south of Brazil. It considers three power plants located in the Grande basin, a part of the Parana basin with two main rivers, the Grande and Pardo rivers. Caconde (CAC) reservoir plant and Euclides da Cunha (EDC) run off hydro plant are part of a cascade at the Pardo river, whereas the Água Vermelha (AVE) reservoir hydro plant is located in the Grande river cascade. All these power plants are under the management of AES-Brasil company.

Data available correspond to a continuous period of three years from 1st January 2005 to 31st December 2007. Accumulated daily rainfall data for the catchment area associated with each plant were supplied by SOMAR company while average daily inflow data were supplied by AES. Table I summarizes some statistics observed during the indicated period for the overall time series where Max = Maximum, Min = Minimum and $S Dev$ = Standard Deviation.

From this table, we can notice that Agua Vermelha inflows present the highest mean and variability, whereas Canconde has the lowest ones. On the other hand, Agua Vermelha presents an average rainfall lower than the ones registered in the other two hydro plants, but both – Caconde and Euclides da Cunha – present similar rainfall statistics in terms of mean and standard deviation, specially in terms of magnitude level if compared to Agua Vermelha.

TABLE I SOME STATISTICS FOR ALL THE TIME SERIES USED IN THE CASE STUDY.

Alias	Inflows				Runoff			
	Max	Min	Mean	S. Dev	Max	Min	Mean	S. Dev
CAC	339	14	59	53	711	0	4.07	9.23
EDC	714	21	95	90	686	0	4.13	9.45
AVE	8641	537	2316	2066	2573	0	3.65	8.38

B. Methodology and Modeling

Input patterns consider both, past and future precipitation information and past inflow data. Therefore, a general input pattern is defined as:

$$x^k = [x_1^k \ x_2^k \ \dots \ x_p^k] = [r_1^k \ \dots \ r_a^k \ s_1^k \ \dots \ s_b^k] \quad (6)$$

where $a \in Z^+$ indicates the number of components of the input vector containing past and future rainfall information represented by r and $b \in Z^+$ represents the number of input vector terms containing past inflow records (s), with $p = a + b$. It means that if we are forecasting the streamflow for time instant k , rainfall terms composing the input patterns will consider the last $a-1$ accumulated D days rainfall lags, the expected accumulated D days rainfall for instant k , as well as the last b inflows recorded up to $k-1$ as input variables. For the next h steps ahead, forecasted inflows and rainfall will be necessary.

With the purpose of improving the explanatory power of rainfall time series for each hydro plant, the FIS model uses as rainfall variable the accumulated precipitation of the last D days, where D is selected by maximizing the correlation coefficients between observed runoff and accumulated precipitation during those D days. Optimal D values for all the three inflow time series as well as correlation coefficient achieved are presented in Table II.

TABLE II OPTIMAL D VALUES, CORRELATION COEFFICIENTS AND LAGS CONSIDERED FOR BUILDING INPUT PATTERNS CONSIDERING RAINFALL AND RUNOFF INFORMATION.

Alias	D	Correlation	a	b
CAC	5	0.73	5	3
EDC	5	0.71	5	3
AVE	11	0.75	4	3

Table II also shows the number of lags a and b considered as input variables for modeling each model. These numbers of lags to were selected by maximizing the model performance for a multi-step ahead daily inflow forecasting task, with time horizon varying from one to twelve days ahead. As observed, in the case of runoff (b), the first three lags of each streamflow time series were considered as input variables. In the case of accumulated precipitation, lags considered varied from 5 to 11 days.

After input-output patterns are built, data set is normalized to the unit interval and then split up in two subsets, a training data set and a testing data set. The training data set was composed by the first two years of historical data whereas the testing data set was composed by the last one. Initial spreads for the SC algorithm (r_a^2) was equal to 1.25.

Performance assessment and model validation was performed considering the root mean square error (RMSE (m^3/s)), mean absolute percentage error (MAPE (%)), mean absolute error (MAE (m^3/s)) and mass curve coefficient (E), which is frequently used in hydrology for the validation of hydrological models [14] and is defined by the following equation:

$$E = \frac{\sum_{k=1}^N (y^k - \bar{y})^2 - \sum_{k=1}^N (y^k - \hat{y}^k)^2}{\sum_{k=1}^N (y^k - \bar{y})^2} \quad (7)$$

where $\bar{y}$ is the observed mean inflow. The first sum in the numerator denotes the total square error (SE) achieved assuming as forecast inflow the historical mean for all k . The second sum denotes the SE achieved by the evaluated model (FIS). Therefore, the closer E is to the unity, the better the model is, since the respective SE will be lower than the SE given by the mean. In this case the use of a more complex model would be justified by the improvement of the accuracy in the forecasting process.

C. Tool Developed

A tool to automate the application of the proposed methodology was developed. The software project and the computational implementation of this tool use the object-oriented paradigm and the C++ programming language. For searching information in the database the Structured Query Language (SQL) is used.

Fig. 2 presents the main interface of the system. It allows managing streamflow forecasting studies. Besides that, it facilitates the analysis of the results because these can be visualized in graphical form.

The interface presented in Fig. 2 is divided in frames. In the first frame on the left, the studies of streamflow forecasting are shown using a hierarchical structure for

visualization. In the next frame, on the middle, the hydroelectric plants and the measurements stations are listed. Different icons distinguish them. The last frame on the right presents five sheets of which three have highlighted. The first one make possible to choose the period for which the study will be realized. So, this tool is able to predict the future streamflows or trying to reproduce the past streamflows. The second sheet shows one chart of recorded and predicted inflows and one chart of the precipitations for the period. The third sheet presents the results in spreadsheets.

This tool allows eight actions with the studies: create, duplicate, rename, delete, save, import data, edit data and run the study. Besides, it is possible export the results of a study for electronic spreadsheets, like Excel.

D. Results and Discussions

Performance metrics are summarized in Table III, for a given forecasting horizon h varying from 1...12. Since the FIS provides a different forecasted trajectory for each time instant, columns 2 to 13 present average errors of the last 353 trajectories predicted from one to twelve steps ahead.

As expected, observing any of the metrics, because of the prediction error of previous steps are fed into the input pattern for the next step ahead, the model performance becomes degraded as the forecast horizon aparts from the actual time instant. However, given that the minimum value of the mass curve coefficient goes from almost 77% to 99%, we can consider that the model presents an adequate performance due to its capability of explaining the hydrological process in at least 77%.

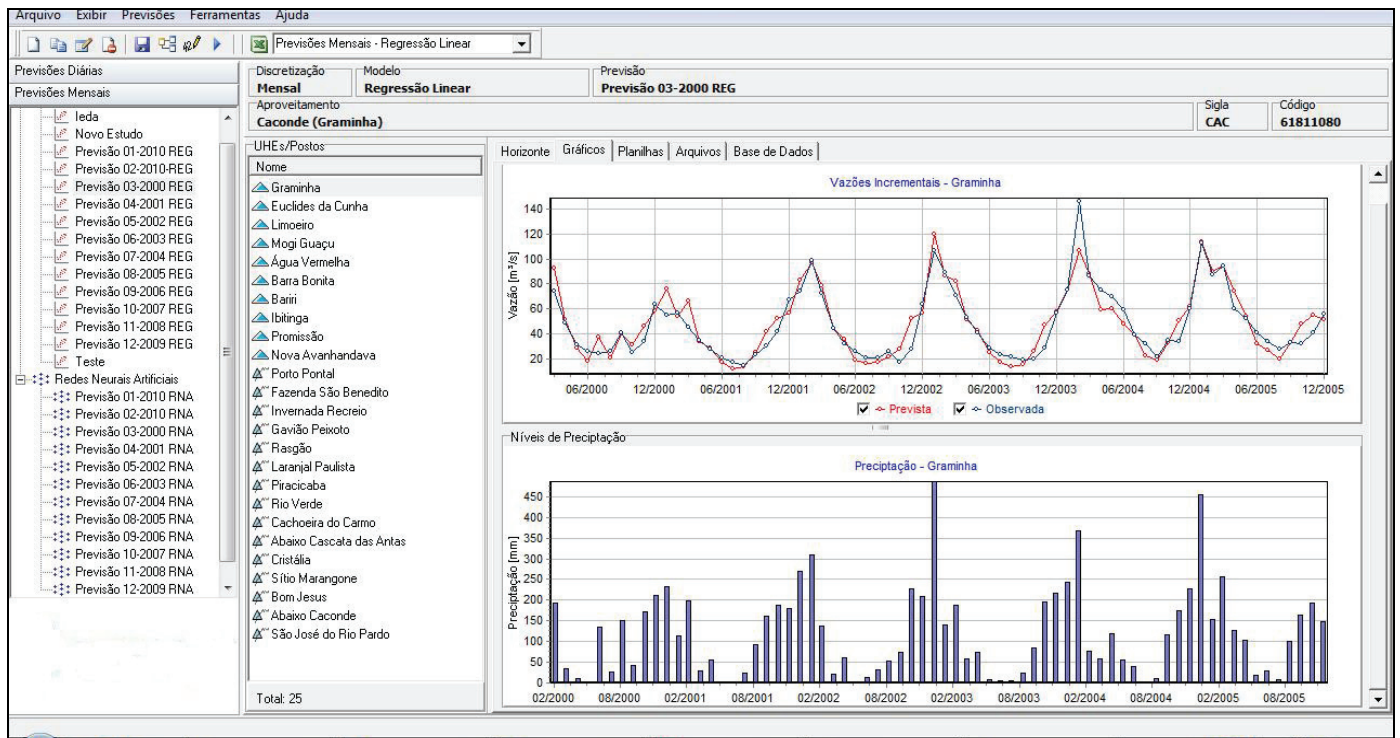


Figure 2. Main interface of the tool developed for forecasting streamflows.

TABLE III PERFORMANCE METRICS.

Alias	Horizon (h)											
	1	2	3	4	5	6	7	8	9	10	11	12
MAPE (%)												
CAC	8.96	11.32	12.91	14.43	15.78	16.91	17.85	18.66	19.36	20.01	20.67	21.29
EDC	6.98	8.91	10.33	11.40	12.27	12.98	13.56	14.01	14.41	14.75	15.05	15.33
AVE	1.62	2.46	3.35	4.20	5.05	5.84	6.61	7.35	8.06	8.74	9.39	10.00
RMSE (m³/s)												
CAC	0.74	0.88	0.97	1.04	1.10	1.15	1.20	1.24	1.27	1.30	1.33	1.35
EDC	1.25	1.55	1.71	1.82	1.90	1.99	2.06	2.13	2.17	2.21	2.25	2.29
AVE	6.16	8.00	10.25	12.20	14.32	16.21	18.07	19.78	21.44	23.01	24.53	25.95
MAE (m³/s)												
CAC	6.04	7.56	8.53	9.37	10.05	10.64	11.11	11.53	11.89	12.23	12.57	12.89
EDC	9.76	11.75	13.29	14.33	15.17	15.86	16.46	16.87	17.24	17.53	17.80	18.07
AVE	44.23	63.97	84.66	103.55	122.54	140.94	158.92	176.22	192.85	208.56	223.22	236.83
E												
CAC	0.94	0.92	0.90	0.88	0.86	0.85	0.83	0.82	0.80	0.79	0.78	0.77
EDC	0.94	0.91	0.88	0.87	0.85	0.84	0.82	0.81	0.80	0.79	0.78	0.77
AVE	0.99	0.99	0.99	0.99	0.98	0.98	0.97	0.97	0.96	0.96	0.95	0.94

Fig. 3 depicts the evolution of MAPE and E from one up to twelve days ahead of forecasted inflow records. Since Caconde and Euclides da Cunha inflows belong to a different magnitude level if compared to Agua Vermelha, their results will be compared and analyzed simultaneously.

The worse performance in terms of MAPE was achieved for Caconde hydroplant. Nevertheless, given that this metric is a relative index, it is necessary to consider at least one other absolute performance metric. So, analyzing RMSE and MAE values, results achieved for Caconde outperformed the ones obtained for Euclides da Cunha. Though, considering the mass curve coefficient, the fuzzy system achieved a similar performance for both inflow time series, as also depicted in Fig. 3.

Forecasted trajectories for one step ahead and for one up to twelve steps ahead are depicted in Fig. 4 and 5, for Caconde

and Euclides da Cunha, respectively. We can observe that, the difficulty for predicting inflows increase during humid periods, since the different trajectories obtained at the beginning of the year tend to apart from the observed one with the increase of the lead time and specially before the peak flows. In general, performance metrics point to an adequate and similar model performance in both cases, which is expected due to the nature of Euclides da Cunha hydroelectric (a run off plant).

On the other hand, considering results obtained for Agua Vermelha, the FIS presented the lowest MAPE but the highest MAE, which is expected due to the inflows levels. In terms of explanatory power, the FIS presented a mass curve coefficient varying from 94% to 99%. Forecasted results are illustrated in Fig. 6. Again, we can notice the effect of the lead time and uncertainties over the forecasted inflow trajectories.

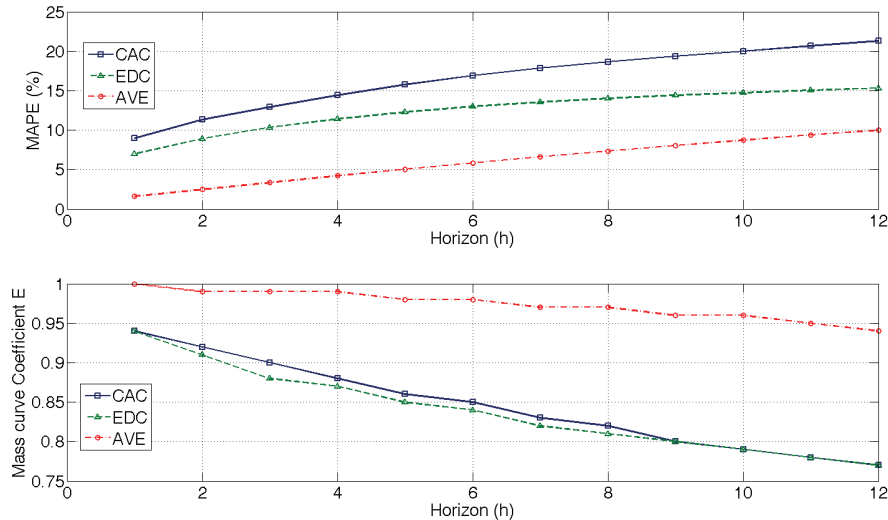


Figure 3. Mean absolute percentage error and mass curve coefficient evolution.

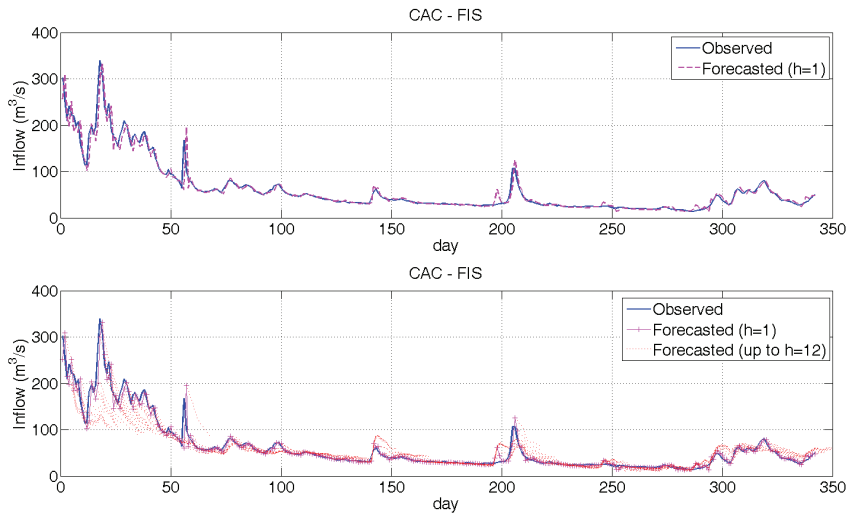


Figure 4. One step ahead and one up to twelve steps ahead daily inflow forecasting for Caconde hydroplant.

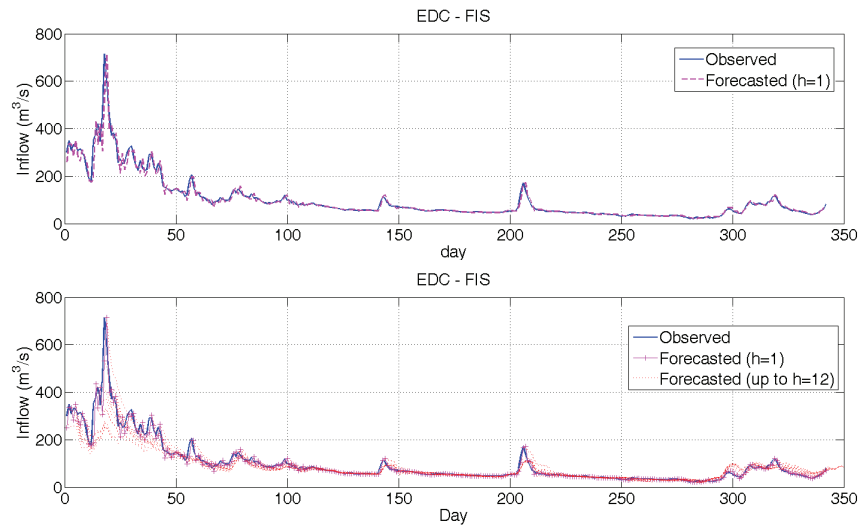


Figure 5. One step ahead and one up to twelve steps ahead daily inflow forecasting for Euclides da Cunha hydroplant.

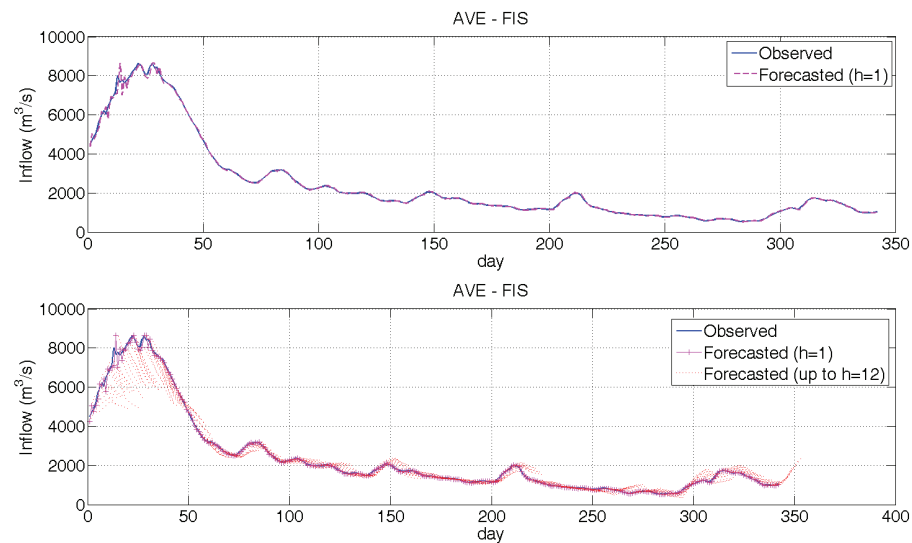


Figure 6. One step ahead and one up to twelve steps ahead daily inflow forecasting for Agua Vermelha hydroplant.

IV. CONCLUSIONS

This work presented an application of Fuzzy Inference Systems (FIS) for multi-step ahead daily inflow time series forecasting. The technique used for adjusting of the model parameters is an offline version of the Expectation Maximization (EM) algorithm.

The case study considered three Brazilian hydroelectric plants: Caconde (CAC), Euclides da Cunha (EDC) and Água Vermelha (AVE) that are located in the Parana basin. Their operation is under the coordination of Electric System National Operator and under management of AES-Brasil company.

A tool was developed in order to automate the application of the methodology and facilitate the analysis of the results. This tool allows managing streamflow forecasting studies and visualizing their information in graphical form.

Although the results shows very promising, it is evident the need research about alternative for improving multi-step ahead forecasting results. Future works involve the use of a different learning algorithm for the fuzzy inference system in order to avoid the normality hypothesis over the time series distribution, as well as the use of adaptive reorganizable fuzzy model in order to deal with the higher variability presented during humid periods. Comparison with other proposals is also considered.

REFERENCES

[1] C. Dawson and R. Wilby, "Hydrological modelling using artificial neural networks," *Progress in Physical Geography*, vol. 25, no. 1, pp. 80–108, 2001.
 [2] Jun Zhang, Chun-tian Cheng, Sheng-li Liao, Xin-yu Wu, and Jian-jian Shen "Daily reservoir inflow forecasting combining QPF into ANNs model". *Hydrol. Earth Syst. Sci. Discuss.*, 6, 121-150, 2009.
 [3] R. Price, "Lecture on knowing the context between hydroinformatics and flood modelling," UNESCO-IHE Institute for Water Edutation, Tech. Rep., November 2008.
 [4] C. M. Zealand, D. H. Burn, and S. P. Simonovic, "Short

term streamflow forecasting using artificial neural networks," *Journal of Hydrology*, no. 214, pp. 32–48, 1999.
 [5] P. Nayak, K. Sudheer, D. Rangan, and K. Ramasastri, "A neuro-fuzzy computing technique for modeling hydrological time series," *Journal of Hydrology*, vol. 291, pp. 52–66, 2004.
 [6] G. J. Bowden, H. R. Maier, and G. C. Dandy, "Input determination for neural network models in water resources applications. Part 2. Case Study: forecasting salinity in a river," *Journal of Hydrology*, no. 301, pp. 93–107, 2005.
 [7] M. Zambelli, I. Luna, and S. Soares, "Long-Term hydropower scheduling based on deterministic nonlinear optimization and annual inflow forecasting models." *Procs. of the PowerTech Conference*, pp. 1–8, 2009.
 [8] I. Luna, S. Soares, J.E.G. Lopes, and R. Ballini. Verifying the use of evolving fuzzy systems for multi-step ahead daily inflow forecasting. 15th International Conference on Intelligent System Applications to Power Systems – ISAP '09, pages 1–6, November 2009.
 [9] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Transactions on Systems, Man and Cybernetics*, no. 1, pp. 116–132, January/February 1985.
 [10] R. Jacobs, M. Jordan, S. Nowlan, and G. Hinton, "Adaptive mixture of local experts." *Neural Computation*, vol. 3, no. 1, pp. 79–87, 1991.
 [11] Lázaro, M., Santamariaa, I. e Pantaleón, C. (2003). A new EM-based training algorithm for RBF networks, *Neural Networks* 16(1): 69–77.
 [12] S. Chiu, "A cluster estimation method with extension to fuzzy model identification," in *Proceedings of The IEEE International Conference on Fuzzy Systems*, vol. 2, June 1994, pp. 1240–1245.
 [13] Ivette Luna, Leandro Maciel, Rodrigo Lanna F. da Silveira, and Rosangela Ballini. Estimating the Brazilian central bank's reaction function by fuzzy inference system. In Eyke Hüllermeier, Rudolf Kruse, and Frank Hoffmann, editors, *IPMU (2)*, volume 81 of *Communications in Computer and Information Science*, pages 324–333. Springer, 2010.
 [14] A. Aitken, "Assessing systematic errors in rainfall-runoff models," *Journal of Hydrology*, vol. 20, pp. 121–136, 1973.