Fuzzy inference systems for synthetic monthly inflow time series generation

Ivette Luna[1] **Rosangela Ballini**[1] **Secundino Soares**[2] **Donato da Silva Filho**[3]

[1]Institute of Economics, UNICAMP, Sao Paulo, Brazil 13083–857, {ivette,ballini}@eco.unicamp.br
[2]School of Electrical and Computer Engineering, UNICAMP, Sao Paulo, Brazil 13083–852, dino@cose.fee.unicamp.br
[3]EDP Bandeirante, Sao Paulo, Brazil, donato.filho@edpbr.com.br

Abstract

Inflow data plays an important role in water and energy resources planning and management. In general, due to the limited availability of historical inflow data, synthetic streamflow time series have been widely used for several applications such as mid- and long-term hydropower scheduling and the identification of hydrological processes. This paper explores the use of fuzzy inference systems for the identification of two hydrological processes, and its use in the generation of synthetic monthly inflow sequences. Experiments using Brazilian monthly records show that fuzzy systems provide a promising approach for synthetic streamflow time series generation.

Keywords: Fuzzy inference systems, synthetic time series, inflow data, stochastic process.

1. Introduction

Inflow time series are an essential component in energy and water resources planning and management. Some of the concerns in hydropower scheduling are the stochastic nature of inflows and the generally limited duration of the historic inflow time series available. In order to improve the description provided by the observed data, synthetic time series are usually used.

Several proposals have been made in the literature for the generation of synthetic series. The most popular model used for modelling hydrological processes on a monthly basis is the autoregressive moving average (ARMA) model [1]. Concerns about this model are related to the determination of adequate data transformation, since ARMA models assume a static nature for the deseasonalized series (stationarity), which contradicts empirical evidence [2].

In order to overcome this problem, different approaches based on computational intelligence models have appeared in recent decades. Such tools are particularly powerful in situations where it is difficult to determine the actual physical process. Artificial neural networks (ANN) are what is most often used for this purpose. The proposal detailed in [3] suggests the ANN as a viable alternative for multivariate generation of monthly inflow series. Furthermore, the work detailed in [4] shows that ANN models are able to generate synthetic inflow series which are statistically similar to those actually observed, outperforming ARMA models.

Fuzzy rule-based systems and fuzzy clustering algorithms are another option which are widely used for inflow forecasting [5], [6], [7]). These papers have concluded that fuzzy models are able to deal with nonlinearities inherent in hydrological processes and that they provide an adequate performance in forecasting tasks.

This paper suggests a fuzzy inference system (FIS) for synthetic monthly inflow generation. The model structure is given by a set of fuzzy rules, which are initialized using a Subtractive Clustering algorithm (SC), originally proposed in [8]. This initialization already provides a FIS with singletons as consequents. Another FIS has been obtained via parameter optimization using the Expectation maximization (EM) algorithm, as detailed in [9].

Inflow innovations are built based on the FIS models for representing the deterministic component, whereas the stochastic one is determined using a bootstrapping technique. Comparison of results obtained show the problem of assuming a normal distribution over observed data when the theoretical distribution is assymetric and unknown, as well as the capability of the FIS models for the generation of synthetic inflow time series.

The rest of the paper is organized as follows. The next section presents the structure of the fuzzy inference system and the learning algorithm involved. The methodology used to generate a synthetic data, as well as an experimental evaluation using real Brazilian inflow series, are detailed in Section 3. Finally, conclusions and suggestions for future research are presented in Section 4.

2. Fuzzy inference system

2.1. Structure

Let $\mathbf{x}^k = [x_1^k,\ x_2^k,\ \ldots,\ x_p^k] \in \mathbb{R}^p$ denote the input vector at instant k, $k \in \mathbb{Z}_0^+$; $\hat{y}^k \in \mathbb{R}$ is the output model, for the correspondent input $\mathbf{x}^k$. The input space represented by $\mathbf{x}^k \in \mathbb{R}^p$, is partitioned into M sub-regions, each represented by a fuzzy rule; $k = 0, 1, 2, \ldots$ is the time index (Figure 1). The antecedents of each fuzzy **If-Then** rule (R_i) are represented by their respective centers $\mathbf{c}_i \in \mathbb{R}^p$ and covariance matrices $\mathbf{V}_i|_{p\times p}$. The consequents are represented by local linear models, with output y_i, $i = 1, \ldots, M$ defined by:

$$y_i^k = \phi^k \times {\theta_i}^T \qquad (1)$$

where $\phi^k = [1\ x_1^k\ x_2^k\ \ldots x_p^k]$; $\theta_\mathbf{i} = [\theta_{i0}\ \theta_{i1}\ \ldots\ \theta_{ip}]$ is the coefficient vector of the local linear model for the i^{th} rule.

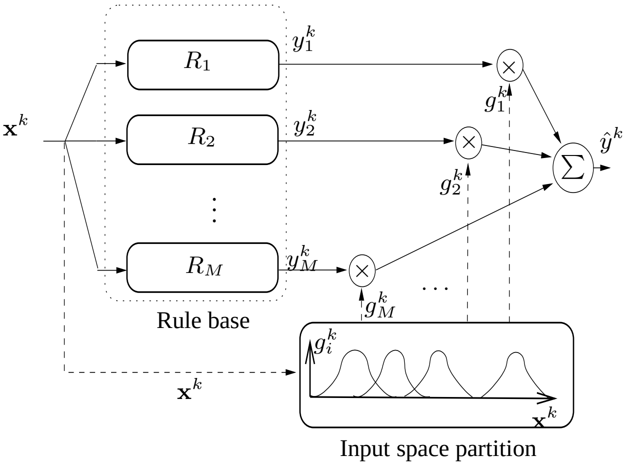

Figure 1: General FIS formulation.

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

$$g_i(\mathbf{x}^k) = g_i^k = \frac{\alpha_i \cdot P[\ i \mid \mathbf{x}^k\ ]}{\sum_{q=1}^{M} \alpha_q \cdot P[\ q \mid \mathbf{x}^k\ ]} \qquad (2)$$

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

$$P[\,i \,|\, \mathbf{x}^k\ ] \quad = \quad \frac{1}{(2\pi)^{p/2}\det(\mathbf{V}_i)^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}^k - \mathbf{c}_i)\mathbf{V}_i^{-1}(\mathbf{x}^k - \mathbf{c}_i)^T\right\} \quad (3)$$

where $\det(\cdot)$ is the determinant function. The model output $y(k) = \hat{y}^k$, which represents the predicted value for future time instant k, is calculated by means of a non-linear weighted averaging of local outputs y_i^k and its respective membership degrees g_i^k, i.e.

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

2.2. Optimization

Model structure is initialized using the Subtractive Clustering Algorithm (SC), an unsupervised clustering algorithm proposed in [8]. This algorithm provides a set of M clusters from a specific training data set presented to the algorithm. Patterns processed by the SC algorithm are composed of the input-output patterns to be used in a second stage for model optimization.

These groups are associated with a set of fuzzy rules codified in the FIS structure. Therefore, after the number of fuzzy rules is defined, we proceed to initialize the model parameters for $i = 1, \ldots, M$, according to the following criteria:

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

This initial structure is a simple fuzzy-rule based system with consequents defined by singletons (FIS-S).

After this initialization, model parameters are readjusted on the basis of the offline EM algorithm (see [9] for the complete formulation), with the objective of maximizing the log-likelihood $\mathcal{L}$ of the observed values of y^k at each step M of the learning process. This objective function is defined by

$$\mathcal{L}(D,\, \mathbf{\Omega}) = \sum_{k=1}^{N} \ln\left(\sum_{i=1}^{M} g_i(\mathbf{x}^k,\ \mathbf{C}) \times P(y^k \mid \mathbf{x}^k,\ \theta_i)\right) \qquad (5)$$

where $D = \{\mathbf{x}^k, y^k | k = 1, \ldots, N\}$; $\mathbf{\Omega}$ contains all model parameters and $\mathbf{C}$ contains just the antecedents parameters (centers and covariance matrices). The FIS model obtained after EM optimization is known as FIS-EM. As observed, for maximizing $\mathcal{L}$, it is necessary to know the data distribution. Since this probability distribution is unknown, the FIS-EM model must be adjusted by assuming a normal distribution for the observed records.

3. Methodology and case study

The FIS-S and FIS-EM models were applied in the generation of 2000 years of synthetic monthly inflows for two plants of the Brazilian hydroelectric system. These plants, the Furnas and Peixoto plants are part of a cascade on the Rio Grande river, located in the southern part of Brasil. Historical time series consist of monthly records from 1931 to 2009. Twelve different models were adjusted, one for each month, since due to wet and dry periods over the year, each month has unique features in terms of statistics and probability distribution.

Input-output data was normalized between 0 and 1. Model selection utilized the past ten years of historical data as a validation dataset, and the model with the lowest Bayes Information Criterion (BIC) [10] was selected as the most adequate for each month. Therefore, the model selection considered the choice of input variables as well as the choice of the spread parameters r_a and r_{ba} (used by the SC algorithm, where r_{ba} represents the distance between centers found by the algorithm). Parameter r_a varied from 0.25 to 1.0 whereas r_{ba} varied from 1.0 to 2.0. The set of possible input variables was defined by the first five lags of the series. All the models selected incorporated a single lag input, except for June and July, where the best configurations consisted of the first two lags.

After model adjustment, we proceeded to the generation of the synthetic series. As mentioned in Section 1, the deterministic component was represented by the FIS model (FIS-S or FIS-EM), while the stochastic portion was defined using a bootstrap resampling technique considering the replacement of the elements resampled.

According to [1], non-parametric techniques such as those based on bootstrapping may capture any distributional information retained in the residuals of a data-driven model. Therefore, residuals estimated from the historic sequence used for FIS adjustment were calculated, and a sample was randomly selected for representing the stochastic part of the simulated innovations. This random selection assumed that residuals were independent and identically distributed (i.i.d.) and following a uniform distribution so that all the residuals for a given model would have the same chance of being selected. As a consequence, although the actual distribution of the series was unknown, it was assumed that the empirical density function of the original observations would be preserved.

Therefore, the final synthetic innovation can be represented as follows:

$$z_m^k = FIS_m(\mathbf{x_m}^k) + \epsilon_m^{k^*} \qquad (6)$$

where z_m^k represents the k^{th} synthetic streamflow for the m^{th} month, FIS_m represents the fuzzy model adjusted for the m^{th} month, $\mathbf{x_m}^k$ is the input vector for the FIS_m used for generating z_m^k and $\epsilon_m^{k^*}$ is the bootstrapped residual selected for building the k^{th} synthetic replicate related to month m. The synthetic series was initialized considering the respective long term historical monthly average as the first twelve innovations, although the first year of the synthetic series was then disregarded to eliminate the effect of the initialization.

The statistics considered to compare the synthetic time series with the historical data were mean monthly value of inflow, monthly standard deviation and skewness and kurtosis coefficients. For graphical analysis, histograms, partial autocorrelation functions and qq-plots were also analized.

3.1. Analysis of results

Figure 2 shows the observed histograms for the Furnas and Peixoto plants.

The use of the FIS-EM reveals an adequate performance for forecasting tasks if one assumes a normal distribution over a set of different applications, including monthly inflow forecasting [11], [5]. However, the normality hypotheses about the inflow distribution affects the FIS performance considerably when used for the generation of synthetic inflows. A normal distribution gives the same chances to very low and very high inflows, although the histograms depicted in Figure 2 show that very high inflows are much less likely than are very low ones. An analysis of monthly distributions shows the same behavior, with a greater skewness during drought periods. Table 1 provides the information about observed and and synthetic mean, standard deviation and skewness and kurtosis coefficients for the Furnas inflow time series.

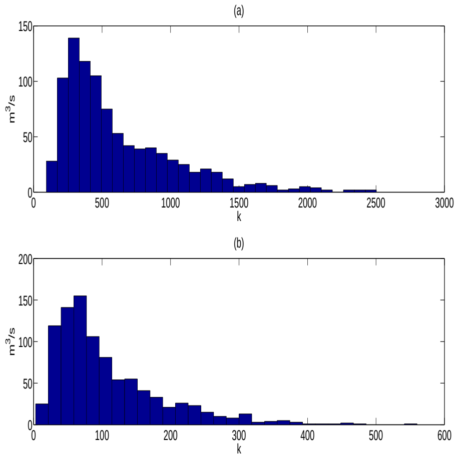

Figure 2: Observed histograms: (a) Furnas, (b) Peixoto.

Figure 3 shows a comparative plot of these summary statistics for the historical and synthetic inflow time series. From these results for the Furnas plant, it can be seen that the assumption of normality considered for the parameter adjustment of the FIS-EM model does not affect its ability to reproduce mean and standard deviation of streamflow series, but it reproduces neither skewness nor kurtosis coefficients. Even though the FIS-S structure is simpler than that of the FIS-EM model, its performance is better in terms of mean and standard deviation; moreover, monthly skewness is better preserved for all of the months except September and October, where both fuzzy models revealed problems. This difficulty can be explained because of deviations during wet periods of some years as observed in the general histogram of monthly inflows depicted in Figure 4-(a). 4 also depicts the synthetic histogram as well as the observed and synthetic autocorrelation function and qq-plots.

In general, the synthetic data resulting by the application of the FIS-S model was able to replicate the observed histogram and preserve the autocorrelation structure, as well as replicating the general statistical characteristics of the time series. However, the qq-plots show that the main difficulty of the FIS-S model is the replication of the highest inflows (peaks) observed during the eighty years of historical data.

The results achieved for the streamflows of the Peixoto plant are summarized in Table 2.

A similar behavior was observed for the two models in relation to the replication of means and standard deviations. However, the FIS-S model outperformed the FIS-EM model in reproducing skewness and kurtosis

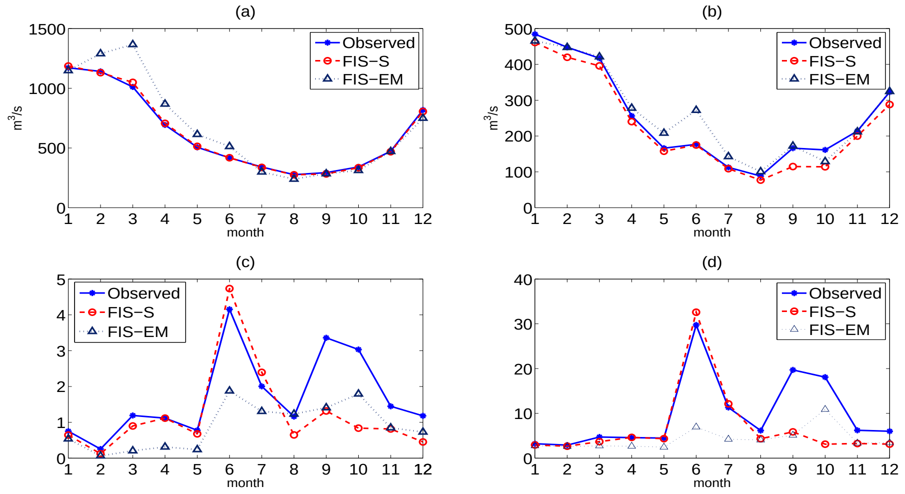

Figure 3: Observed and estimated statistics for the Furnas plant inflow time series: (a) mean, (b) standard deviation, (c) skewness coefficient, (d) kurtosis coefficient.

Table 1: Observed and synthetic statistics for Furnas inflow time series.

Month	1	2	3	4	5	6	7	8	9	10	11	12
	Mean											
Observed	1171	1141	1011	693	506	418	341	277	294	341	473	818
FIS-S	1186	1131	1049	706	514	418	338	276	283	335	469	806
FIS-EM	1148	1290	1366	868	614	513	299	240	285	312	472	750
	Standard deviation											
Observed	484	448	417	256	166	177	113	88	166	161	214	323
FIS-S	461	420	396	240	158	175	109	77	115	114	200	288
FIS-EM	465	447	421	278	208	272	143	101	172	129	212	325
	Skewness											
Observed	0.75	0.25	1.19	1.12	0.78	4.16	2.00	1.16	3.36	3.03	1.45	1.18
FIS-S	0.65	0.12	0.90	1.12	0.68	4.73	2.40	0.65	1.32	0.84	0.81	0.45
FIS-EM	0.54	0.06	0.21	0.31	0.24	1.88	1.31	1.23	1.41	1.79	0.85	0.73
	Kurtosis											
Observed	3.21	2.91	4.72	4.58	4.48	29.73	11.31	6.17	19.71	18.09	6.22	6.00
FIS-S	2.93	2.67	3.68	4.63	4.35	32.59	12.11	4.34	5.86	3.12	3.27	3.13
FIS-EM	2.76	2.59	2.74	2.66	2.46	6.95	4.22	4.06	5.14	10.89	3.36	3.43

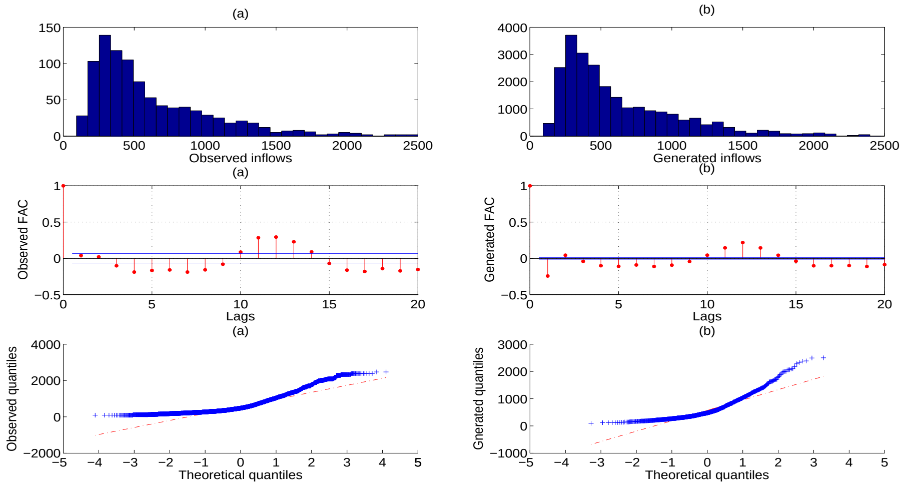

Figure 4: Histogram, autocorrelation function (AFC) and qq-plot of observed and synthetic time series for Furnas: (a) observed, (b) generated.

Table 2: Observed and synthetic statistics for inflow time series of Peixoto plant.

Month	1	2	3	4	5	6	7	8	9	10	11	12
	Mean											
Observed	204	208	186	128	92	72	59	49	48	56	75	128
FIS-S	202	203	185	126	93	76	65	54	49	57	78	129
FIS-EM	201	214	194	131	103	78	73	65	54	58	72	124
	Standard deviation											
Observed	92	99	83	56	35	29	24	21	22	29	38	52
FIS-S	89	95	85	55	33	26	22	18	19	26	36	50
FIS-EM	91	97	82	60	40	34	29	25	21	30	34	54
	Skewness											
Observed	0.82	0.65	0.93	0.43	0.60	0.41	0.70	0.55	0.87	1.29	1.29	0.84
FIS-S	0.65	0.58	0.96	0.40	0.37	0.28	0.46	0.43	0.87	1.45	1.41	0.65
FIS-EM	0.53	0.51	0.64	0.23	0.15	0.15	0.05	-0.08	0.35	0.74	0.85	0.70
	Kurtosis											
Observed	3.81	3.65	3.88	2.69	3.22	2.91	3.67	3.65	3.40	5.21	4.72	3.92
FIS-S	3.66	3.41	3.76	2.58	2.92	2.82	3.29	3.59	3.31	5.55	5.24	3.40
FIS-EM	3.37	2.93	3.10	2.45	2.20	2.09	2.03	2.13	2.69	3.37	3.48	3.14

features for all the month of the year. To facilitate the comparison of the results, statistics calculated for the historical and synthetic series are depicted in Figure 5. The observed and synthetic histograms, autocorrelation function and qq-plots are illustrated in Figure 6.

The figures presented here show the need for consideration of an adequate marginal distribution for the generation of statistically similar synthetic series. Although lack of knowledge about theoretical distribution or an inadequate hypothesis apparently does not affect mean and variance estimations, the reproduction of extreme samples represented by the asymmetric tails of the histogram will not be replicated.

4. Conclusions and suggestions for future work

Preliminary results presented in this paper show fuzzy systems to be a potential tool for the generation of synthetic inflow time series. In general, the means a standard deviations for all months were adequately replicated. On the other hand, the model encountered some difficulties in replicating skewness and kurtosis coefficients for some of the months of streamflow series of the Furnas plant. In general, the data-driven model that was optimized disregarding hypotheses about the marginal distribution of the series outperformed the one that considered a normal data distribution, such as is done by most of the models using the EM algorithm to adjust model parameters. Even though expected means and standard deviations are less affected by this hypothesis, the results show its relevance and its effect on the replication of other statistical aspects such as the skewness and kurtosis coefficients. Therefore, despite the simplicity of the FIS-S model, it was able to provide a reasonable reproduction of summary statistics and marginal distributions.

For future research, the authors intend to develop comparative studies with other models found in the literature, as well as developing statistical tests for the validation of these synthetic series, and the analysis of other hydrological features, including annual distribution and correlation.

5. Acknowledgements

The authors would like to acknowledge the financial support of all the companies of the Brazilian Electrical Sector involved in the R&D project entitled "Optimization Model of Hydrothermal Dispatch", ANEEL code PE-0391-0108/2009, and the Brazilian National Research Council, CNPq.

References

[1] Srinivasan K. Neelakantan T. R. Sudheer, K. P. and V. V. Srinivas. A nonlinear data-driven model for synthetic generation of annual streamflows. *Hydrological Processes*, 22:1831–1845, 2008.

[2] J. R. Stedinger and M. R. Taylor. Synthetic streamflow generation: 1. model verification and validation. *Water Resour. Res.*, 18(4):909–918, 1982.

[3] R. García-Bartual J.C. Ochoa-Rivera and J. Andreu. Multivariate synthetic streamflow generation using a hybrid model based on artificial neural networks. *Hydrology and Earth System Science*, 6(4), 2002.

[4] Juran Ahmed and Arup Sarma. Artificial neural network model for synthetic streamflow generation. *Water Resources Management*, 21:1015–1029, 2007.

[5] 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.

[6] Alexandre Evsukoff, Beatriz Lima, and Nelson Ebecken. Long-term runoff modeling using rain-

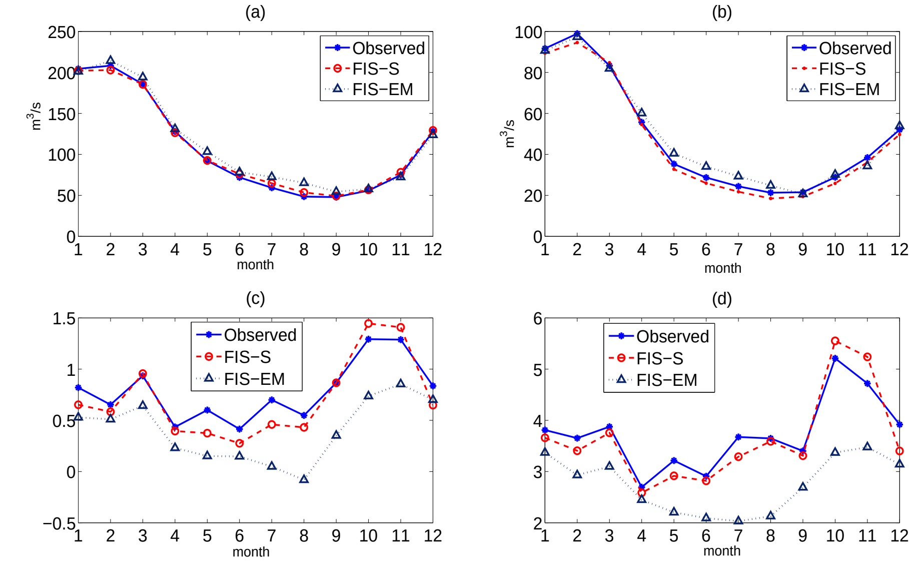

Figure 5: Observed and estimated statistics for inflow time series of the Peixoto plant: (a) mean, (b) standard deviation, (c) skewness coefficient, (d) kurtosis coefficient.

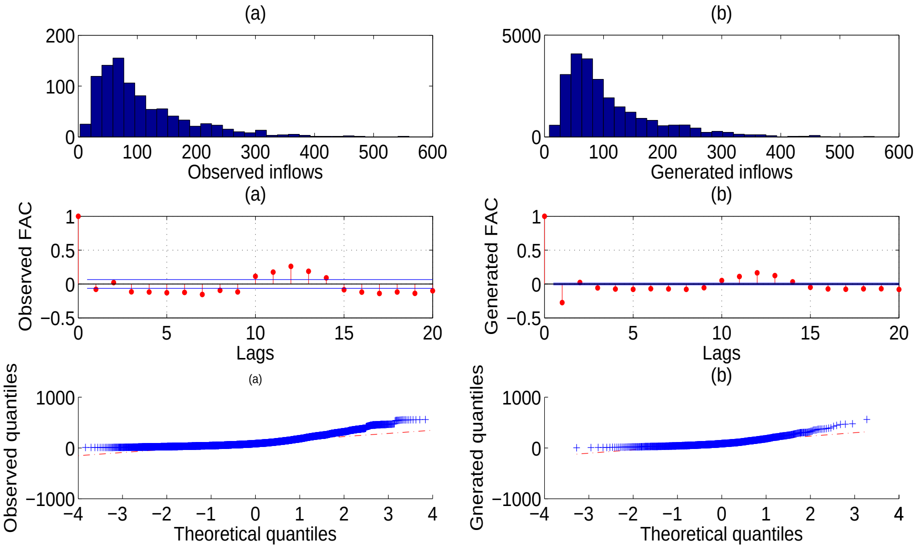

Figure 6: Histogram, autocorrelation function (acf) and qq-plot of observed and synthetic time series for the Peixoto plant: (a) observed, (b) generated.

fall forecasts with application to the iguaçu river basin. *Water Resources Management*, pages 1–23, 2010.

[7] Mahmood Akbari, Peter Overloop, and Abbas Afshar. Clustered k nearest neighbor algorithm for daily inflow forecasting. *Water Resources Management*, pages 1–17, 2010.

[8] S.L. Chiu. A cluster estimation method with extension to fuzzy model identification. In *Proceedings of the Third IEEE Conference on Fuzzy Systems*, volume 2, pages 1240–1245, Orlando - Forida, USA, Junho 1994.

[9] 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.

[10] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–468, 1978.

[11] M.S. Zambelli, I. Luna, and S. Soares. Long-term hydropower scheduling based on deterministic nonlinear optimization and annual inflow forecasting models. In *PowerTech, 2009 IEEE Bucharest*, pages 1–8, Julho 2009.