 New Approaches to Hydrological Prediction in Data-sparse Regions (Proc. of Symposium HS.2 at the Joint IAHS & IAH Convention, Hyderabad, India, September 2009). IAHS Publ. 333, 2009.

Drought forecast using an artificial neural network for three hydrological zones in San Francisco River basin, Brazil

CELSO AUGUSTO G. SANTOS, BRUNO S. MORAIS & GUSTAVO B. L. SILVA
Federal University of Paraíba, Department of Civil and Environmental Engineering, 58051-900 João Pessoa, Brazil
celso@ct.ufpb.br

Abstract Three homogeneous rainfall areas were identified within San Francisco River basin, located in Northeast Brazil, by analysing the rainfall frequencies through the global wavelet power spectra that provide an unbiased and consistent estimation of the true power spectrum of the time series. Such study was accomplished using data from 248 raingauges provided by the Brazil National Water Agency (ANA), for several years between 1938 and 2005, based on their geographical distribution. For each identified region, the standardized precipitation index (SPI) was forecast using a feed-forward artificial neural network (ANN) trained by the back-propagation algorithm. The results obtained show that: the ANN is a suitable tool for this type of forecast; the accuracy is improved when the time scales of the SPI index, as well as the lead times, are increased; and the final result was not influenced by the different hydrological zones.

Key words wavelet; fuzzy; ANN; drought; hydrological zones

INTRODUCTION

About 50% of the more populous areas of the world are highly vulnerable to the drought. In 1967 and 1992, droughts affected 50% of the 2.8 billion people affected by all natural disasters. As a function of the direct and indirect impacts of this phenomenon, 1.3 million human lives were lost, of a total of 3.5 million people who died due to natural disasters (Mishra & Desai, 2006).

In the USA, the current annual costs of droughts are greater than those of any other natural disaster. They are estimated at about US\$6–8 $\times$ 10^9, distributed among agriculture sectors, transportation, tourism and energy (NDMC, 2007). In Brazil, historical records of severe droughts can be observed, mostly, in the semi-arid region of northeastern states: Piauí, Ceará, Rio Grande do Norte, Paraíba, Pernambuco, Alagoas, Sergipe, Bahia, and in northern Minas Gerais. In order to reduce the population vulnerability to drought occurrence, meteorologists and hydrologists have been developing indices based on hydrometeorological variables, which are able to classify droughts in terms of their intensity. In general, these indices are used for drought analysis at different time scales.

McKee *et al*. (1993) developed the Standardized Precipitation Index (SPI) based on precipitation deficit or excess quantification at different temporal scales (1, 2, 3, 6, …, 12 months), that reflect the drought impact in regions of different water availability (McKee *et al*., 1995). The SPI has been used worldwide (Guttman, 1999; Szalai & Szinell, 2000; Lloyd-Hughes & Saunders, 2002; Giddings *et al.*, 2005; Wu *et al.*, 2005), and in Brazil, it is one of the methods recommended by the National Institute of Meteorology (INMET) to characterize precipitation anomalies. Since January 2002, the SPI has been calculated for the whole country and, recently, these results have been made available for the public in a map format at www.inmet.gov.br.

Some factors have contributed to the preference for using the SPI over other indices; for example, the SPI can be determined based on precipitation data only, so drought evaluation is possible even though other hydrometeorological variables are not available; its ease of computation; the versatility in quantifying the precipitation deficit at different time scales; the possibility to compare SPI values for different regions because it is a standardized index (Kim *et al.*, 2006); and the fact that drought is evaluated for different time scales.

In addition to monitoring, the planning actions required to minimize drought effects require the forecast to have a certain amount of lead time. Nowadays, the development of appropriate tools for drought forecasting and warning remains a challenge. Traditionally, stochastic models were used to forecast drought based on temporal series methods. However, these models are basically

linear and have a limited capacity to capture the nonlinearity which is inherent in hydrological phenomena.

In recent decades, the ANN has been demonstrated to have a great capacity in modelling hydrological temporal series, proving to be a useful tool for forecasting natural phenomena. One of the aspects that has motivated the frequent use of ANNs in several scientific fields, particularly in forecasting studies of temporal series, is its proven capacity to adequately represent nonlinear variables.

Based on the reported studies, the SPI values for multiple time scales are computed for a semi-arid Brazilian watershed in order to evaluate the potential of ANNs in drought modelling, and to forecast using the SPI index. Moreover, using the global wavelet power spectra, hydrological zones are determined in the basin, in order to evaluate the SPI forecast quality within distinct regions.

MATERIALS AND METHODS

Location of the study area

The study area is San Francisco River basin, Brazil, which extends from the Serra da Canastra National Park, starting in São Roque de Minas city, to its outlet, between Alagoas and Sergipe states, over approximately 2700 km. San Francisco River basin occupies an area of 364 000 km^2, between 7°00′ and 21°00′S and 35°00′ and 47°40′W. It has 10 sub-basins, which are numbered from 40 to 49.

Selected data

We selected 248 raingauges provided by the Brazil National Water Agency (ANA), with data from several years between 1938 and 2005, based on their geographical distribution. Data of these raingauges were used for the determination of hydrological zones using global wavelet power spectra. For each zone found, some raingauges were selected in order to compute the SPI index for different time scales and to evaluate the forecast quality of this index using ANN within different hydrological zones. Hydrological regionalization is a necessary tool in the basin, in order to help the decision-making process. However, regionalization is complicated because the rainfall, although within the same basin, can present different characteristics, which are not easily detectable due to the similarity of the hydrological regime. For the application of the selected ANN, we used data between the years 1971 and 2005, in which the period 1971–2000 was used for the calibration process and 2001–2005 for the forecasting.

Standardized precipitation index (SPI)

The value of SPI index (McKee *et al.*, 1993) represents the number of standard deviations for the accumulated precipitation, for a specific scale of time (following a gamma probability distribution transformed to a normal distribution), that an event is above or below the mean.

The nature of the SPI allows an analyst to determine the rarity of a drought or an anomalously wet event at a particular time scale for any location in the world that has a precipitation record. A drought event occurs at the time when the value of SPI is continuously negative. The event ends when the SPI becomes positive. Table 1 provides a drought classification based on SPI.

The SPI is computed by fitting a probability density function to the frequency distribution of precipitation summed over the time scale of interest. Typically the time scales used are 3, 6, 9, 12 or 24 months. This is performed separately for each month. Each probability density function is then transformed in to the standardized normal distribution. Once the relationship of probability to precipitation is established from historic records, the probability of any observed precipitation data point is calculated and used along with an estimate of the inverse normal to calculate the precipitation deviation for a normally distributed probability with mean of zero and standard deviation of unity. This value is the SPI for the particular precipitation data point.

Table 1 Classification of the SPI according to McKee *et al*. (1993).

SPI values	Category
2 and above	Extremely wet
1.5 to 1.99	Very wet
1.0 to 1.49	Moderately wet
–0.99 to 0.99	Near normal
–1.0 to –1.49	Moderately dry
–1.5 to –1.99	Severely dry
–2 and below	Extremely dry

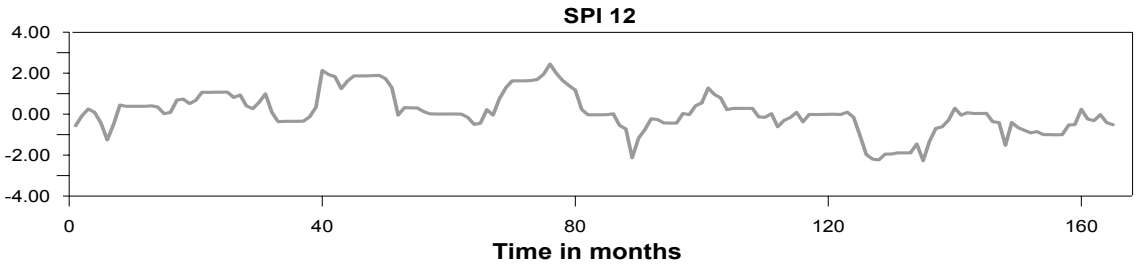

Fig. 1 SPI time series for 12 months time scale (SPI-12).

In order to compute the SPI series, a computational program was developed in a MatLab environment, based on the drought index references, e.g. Mckee *et al*. (1993, 1995). Figure 1 illustrates the computed SPI-12 time series.

Wavelet transform

Wavelet analysis maintains time and frequency localization in a signal analysis by decomposing or transforming a one-dimensional time series into a diffuse two-dimensional time-frequency image simultaneously. Then, it is possible to get information on both the amplitude of any "periodic" signals within the series, and how this amplitude varies with time.

The wavelet analysis is based on a study of the signal convolution $f(t)$ with successive functions, representative of different scales, the wavelet functions $g_{ab}(t)$. The shape of each of these functions is obtained from a primary function, previously defined, commonly called mother wavelet $g(t)$:

$$g_{ab}(t)=\frac{1}{\sqrt{a}}g\left(\frac{t-b}{a}\right) \tag{1}$$

where a (always >0) is the time scale (a smaller scale corresponds to a higher frequency) and b corresponds to the different moments over time. Thus, the wavelet transform (W) is defined as:

$$W(b,a)=\frac{1}{\sqrt{a}}\int g\left(\frac{t-b}{a}\right)f(t)\mathrm{d}t \tag{2}$$

in which the denominator term $\sqrt{a}$ is an energy normalization factor of each wavelet $W(b,a)$ in order to maintain the same energy of the main wavelet. Equation (2) reflects the main innovation in the wavelets, the possibility of transforming a time series in a space of two parameters (a,b) that reflects the local measure on the scale of variability with the scale at the moment b. This definition differs from the equivalent definition of the Fourier transform that gives us only an average range for each scale (frequency or period) of the variability throughout the area. Therefore, two very different signals can have a very similar power spectrum dominated by the same peaks. That is, a simple signal that changes often in the middle of the series and another signal superimposing two frequencies throughout the series would have a power spectrum with two frequencies. Thus,

without any additional knowledge in advance, it would be impossible to affirm which of the two signals produced each of the power spectra. Indeed, all the information on the temporal evolution of the signal is lost when applied the Fourier analysis, which is not the case when applying the wavelet analysis. These limitations have a strong incentive for the development of analysis by wavelets. On the other hand, the global wavelet spectra provide an unbiased and consistent estimation of the true power spectrum of the time series, and thus they are a simple and robust way to characterize the time series variability.

Artificial neural networks

An artificial neural network (ANN) is a mathematical model based on biological neural networks. It consists of an interconnected group of artificial neurons, and processes information using a connectionist approach to computation. In most cases, an ANN is an adaptive system that changes its structure based on external or internal information which flows through the network during the learning phase. In more practical terms, neural networks are nonlinear statistical data modelling tools. Thus, they can be used to model complex relationships between inputs and outputs.

Although, several ANN models have been proposed, the most popular for time series forecasting is the multi-layer feed-forward network. One of the reasons for that is its capacity of universal approach and its flexibility to solve great class problems, including standard recognition, signal processing, control and optimization, classification and forecast problems of time series, although a more robust optimization technique, such as the genetic algorithm (Santos *et al.*, 2003), can be used.

ANN application basically consists of three steps: network architecture definition, network learning, and network verification. The architecture is defined basically by the number of layers, number of neurons per layer, type of connection between layers (activation functions), and the type of network. The learning process consists of supplying the network with an example set and changing their weights until the network is able to represent well the relationship between input and output data. One of the most commonly-used algorithms for ANN learning is the error back-propagation algorithm, which is applied in the present paper. Finally, in the verification process, the network is simulated for a data set which was not used during the learning process. The results are then compared with the observed data in order to verify whether the trained network is able to generalize the results obtained during the learning process.

Developed forecasting model

An ANN multi-layer feed-forward network was built to provide p future values (SPI_{t+1} SPI_t, SPI_{t+2}, ..., SPI_{t+p}) based on previous values (SPI_t, SPI_{t-1} , ..., SPI_{t-m}), in which t represents the current time. In the proposed ANN, only one hidden layer was considered, with a tangent sigmoid activation function between the input and hidden layers, and a linear activation function between hidden and output layers. Figure 2 shows a typical layout of the proposed ANN to a forecast horizon of two months, using three values of SPI as previous input.

The number of neurons in the input (m + 1) and hidden layers was defined by means of previous tests, when the number of neurons was gradually increased in each layer and the obtained error was observed in the learning process for each configuration. Since the smaller error was observed with three neurons in the input layer and 15 neurons in the hidden layer, this architecture was adopted for the present study. No improvement in the results was observed when the number of neurons was larger than this.

For the SPI index forecast, an output neuron quantity (p) is used corresponding to the number of months it was required to forecast, in which the last output neuron was the forecast value for the desired month horizon. Thus, a model was built for each evaluated ANN forecasting horizon. In this type of strategy, used in forecasting and referred to as the direct approach, the values of the variable under study are expected for p steps ahead. The advantage of this strategy is that the errors of predicted values are not accumulated for the next forecast.

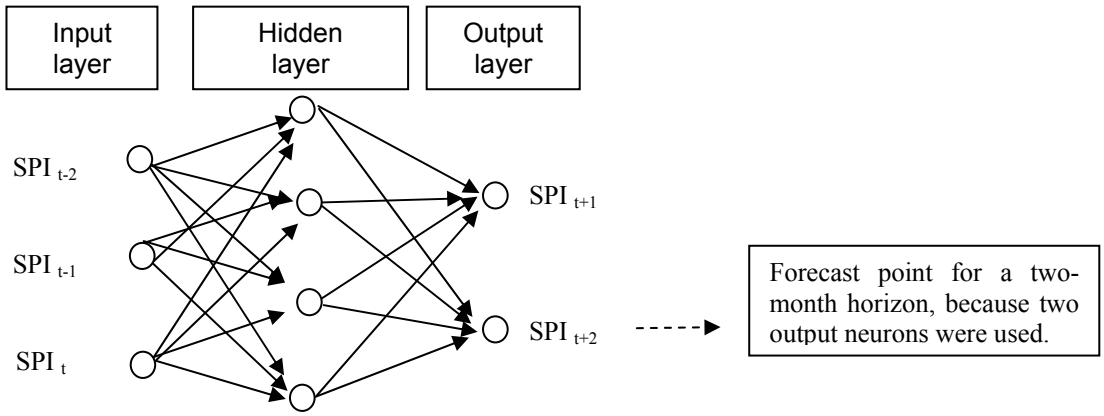

Fig. 2 ANN model in which the output neurons quantity corresponds to the wished forecast month quantity, in which the last output neuron is the forecasted value for the wished month horizon.

For the computed deviation in each forecast, the mean square error (E) was used, as calculated by:

$$E = \frac{\sum_{i=1}^{n} (\mathrm{SPI}_i - \hat{\mathrm{SPI}}_i)^2}{n} \tag{3}$$

in which SPI_i is the SPI value, $\hat{\mathrm{SPI}}_i$ is the forecast value and n is the number of forecast points.

RESULTS AND DISCUSSION

Determination of the hydrological zones

For the studied basin, the results are similar of those of Fig. 3, which depicts the case of raingauge no. 1845004, located at sub-basin 40. Figure 3(a) shows the raw data for the precipitation, Fig. 3(b) shows the wavelet power spectrum, Fig. 3(c) shows the scale-average wavelet power over the 8 to 16-month band, and Fig. 3(d) shows the global wavelet power spectrum.

Figure 3(b) shows the power (absolute value squared) of the wavelet transform for the monthly rainfall at raingauge no. 1845004 presented in Fig. 3(a), which is a record from 1975 to 2005. The (absolute value)^2 gives information on the relative power at a certain scale and a certain time. This figure shows the actual oscillations of the individual wavelets, rather than just their magnitude. Observing Fig. 3(b), it is clear that there is more concentration of power between the 8 to 16-month band, which shows that this time series has a strong annual signal. The variance of power in the 8 to 16-month band (also confirmed later by Fig. 3(c)) also shows the dry and wet years; i.e. when the power decreases substantially in this band, it means a dry year and when the power is maximum it is a wet year. For example, a dry period can be identified between 1984 and 1990 followed by a wet period until the beginning of 1992. An extreme reduction in power can also be found between the years 1992 and 1999, which corresponds to a dry year followed by a wet period until 2004.

The global wavelet spectra provide an unbiased and consistent estimation of the true power spectrum of the time series. For example, one peak well above the others can be seen in Figs 3(d) and 4(a), which indicates that there is an annual frequency, but other raingauges have two or more peaks instead of one as shown in Fig. 4(b). Following the analogy to the other selected raingauges within the basin, it was possible to identify the two main patterns of global wavelet power spectra: Pattern A with one main annual frequency (Fig. 5(a), (c) and (e)) and Pattern B with more than one main frequency (Fig. 5(b), (d) and (f)). From which three distinct hydrological zones (referred to herein as Region A, Region B and transition zone) were identified, as shown in Fig. 6.

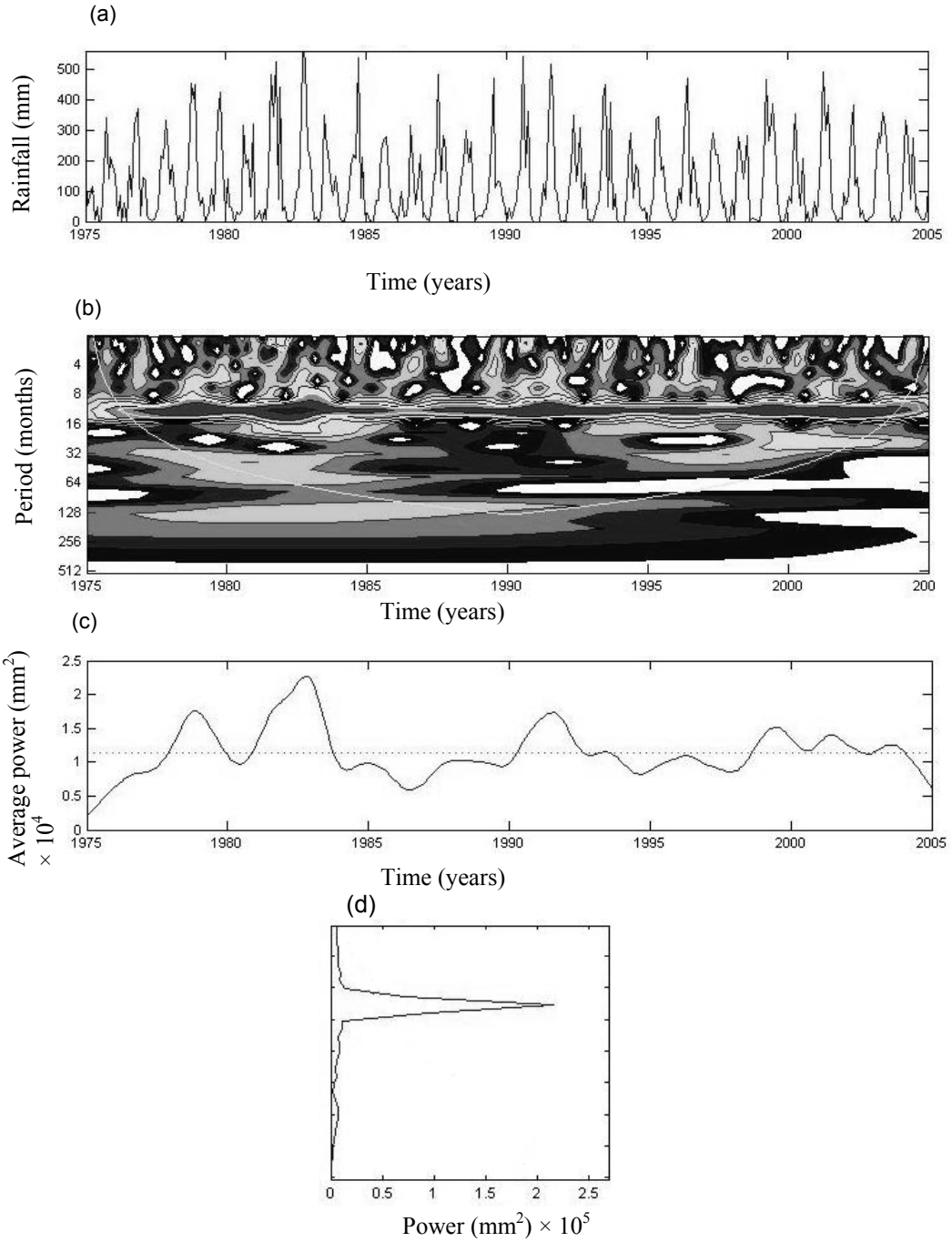

Fig. 3 (a) Monthly rainfall at raingauge no. 1845004 for the 1975–2005 period. (b) The wavelet power spectrum using Morlet mother wavelet. The contour levels are chosen so that 75, 50, 25 and 5% of the wavelet power is above each level, respectively. (c) Scale-average wavelet power over the 8–16-month band. The dashed line is the 90% confidence level assuming red-noise. (d) The global wavelet power spectrum. The dashed line is the 10% significance level for the global wavelet spectrum, using a red-noise background spectrum.

Forecast of the SPI

After determination of the three hydrological zones using the wavelet global spectra, the SPI time series were computed for the scales of 3, 6, 9 and 12 months, in each hydrological zone. Following the SPI time series calculation, using the described ANN, the SPI was forecasted for a horizon of 1 to 6 months. The mean square errors obtained in the forecasts are shown in Table 2.

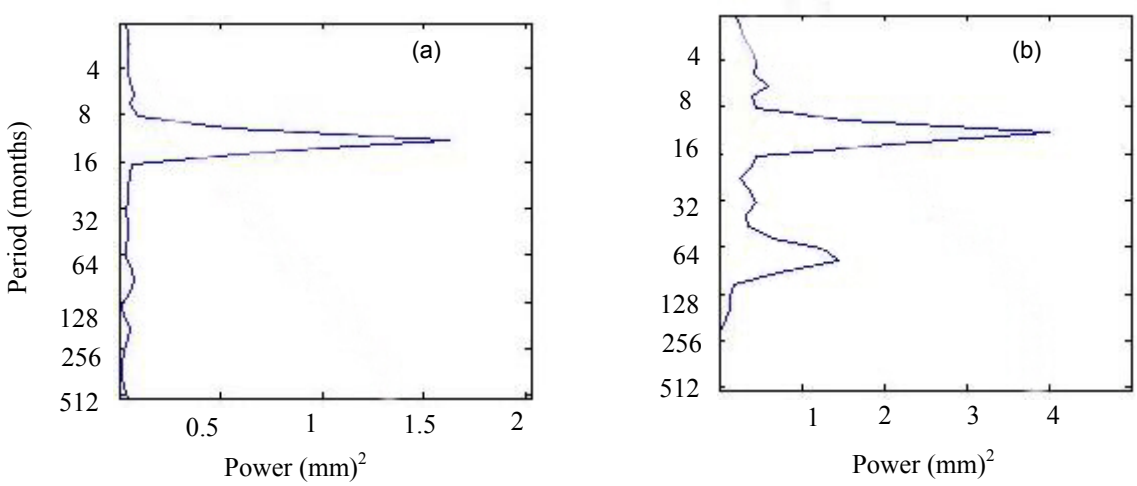

Fig. 4 (a) Global wavelet power spectrum Pattern A, which is characterized by one main frequency at 8–16 months. (b) Global wavelet power spectrum Pattern B, which is characterized by two or more main frequencies, e.g. at 8–16 months and 64–128 months.

Table 2 Mean square error for the forecast of SPI in the three defined hydrological zones.

	1 month	2 months	3 months	4 months	5 months	6 months
Region A:						
SPI-3	0.385 059	0.534 689	0.697 200	1.031 891	0.852 789	1.133 691
SPI-6	0.177 593	0.232 635	0.463 801	0.483 741	0.741 133	0.695 334
SPI-9	0.230 009	0.408 880	0.226 241	0.340 322	0.592 321	0.762 938
SPI-12	0.107 028	0.271 423	0.250 424	0.228 107	0.293 423	0.261 219
Transition zone:						
SPI-3	0.695 870	1.948 702	1.948 702	1.611 778	1.481 205	1.326 161
SPI-6	0.454 619	0.816 986	1.378 087	1.538 248	1.828 336	1.422 181
SPI-9	0.324 792	0.609 811	0.825 017	1.069 271	0.964 212	1.019 718
SPI-12	0.113 656	0.140 329	0.372 329	0.438 147	0.612 805	0.629 826
Region B:						
SPI-3	0.587 873	1.282 542	1.684 702	1.933 310	1.820 806	1.224 206
SPI-6	0.513 189	0.553 021	1.306 867	4.162 345	1.893 579	4.454 624
SPI-9	0.319 986	0.375 936	1.010 807	0.853 261	0.763 810	1.384 081
SPI-12	0.187 253	0.227 718	0.432 096	0.484 548	0.516 106	0.386 680

Analysing the obtained errors, it is possible to see that the result precision increases as the SPI time scale increases. It may be seen that the forecast effectiveness is lower in SPI-3, reaching the best result in the forecast for SPI-12. This is due to the high temporal variability in precipitation in SPI-3, while for the other scales, this variability is attenuated because more monthly data could be collected. Thus, while the SPI time scale increases, the SPI forecast is improved. It is also possible to see that, when the month horizon forecast is increased, a significant increase in the mean square error occurs, which is very common in forecasting models. In Figs 7 and 8 the forecast SPI for the 9- and 12-month time scales are shown, for 1 to 2 months in the three studied regions. It can be observed that there are no significant differences in the forecast model performance when the hydrological zone is changed, which shows that the forecast SPI using the proposed ANN is not strongly affected by the rainfall regime of the region.

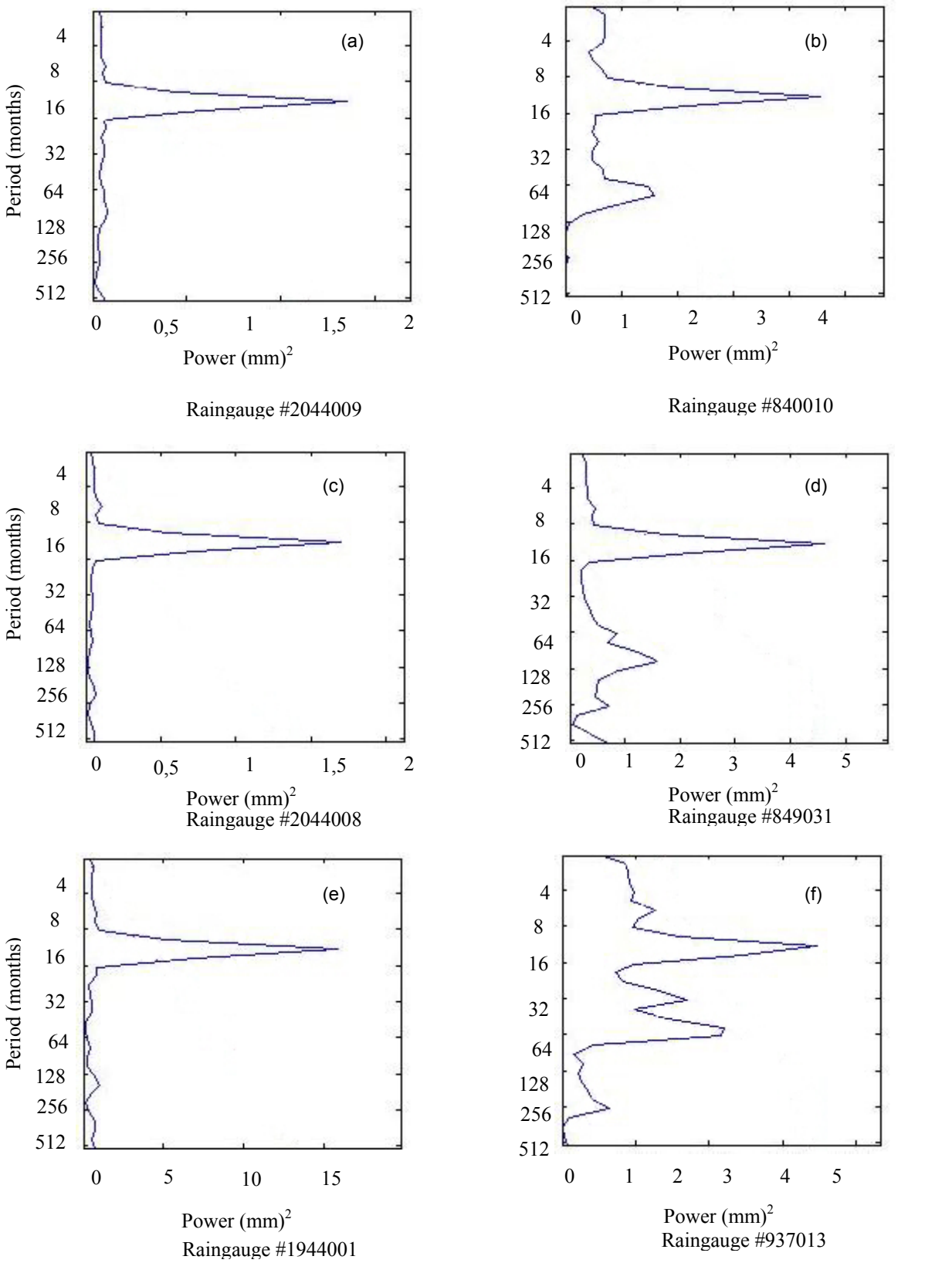

Fig. 5 Global wavelet power spectra: (a,) (c) and (e) for Region A, located in the southern part; and (b), (d) and (f) for Region B, located in the northern part.

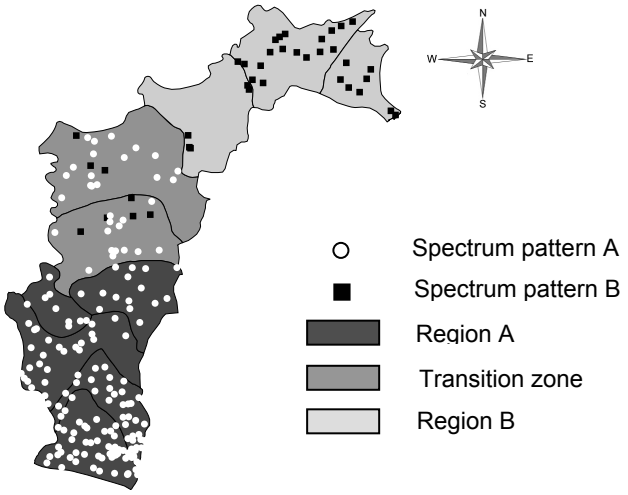

Fig. 6 The raingauges within San Francisco River basin region, according to the pattern of their global wavelet power spectra.

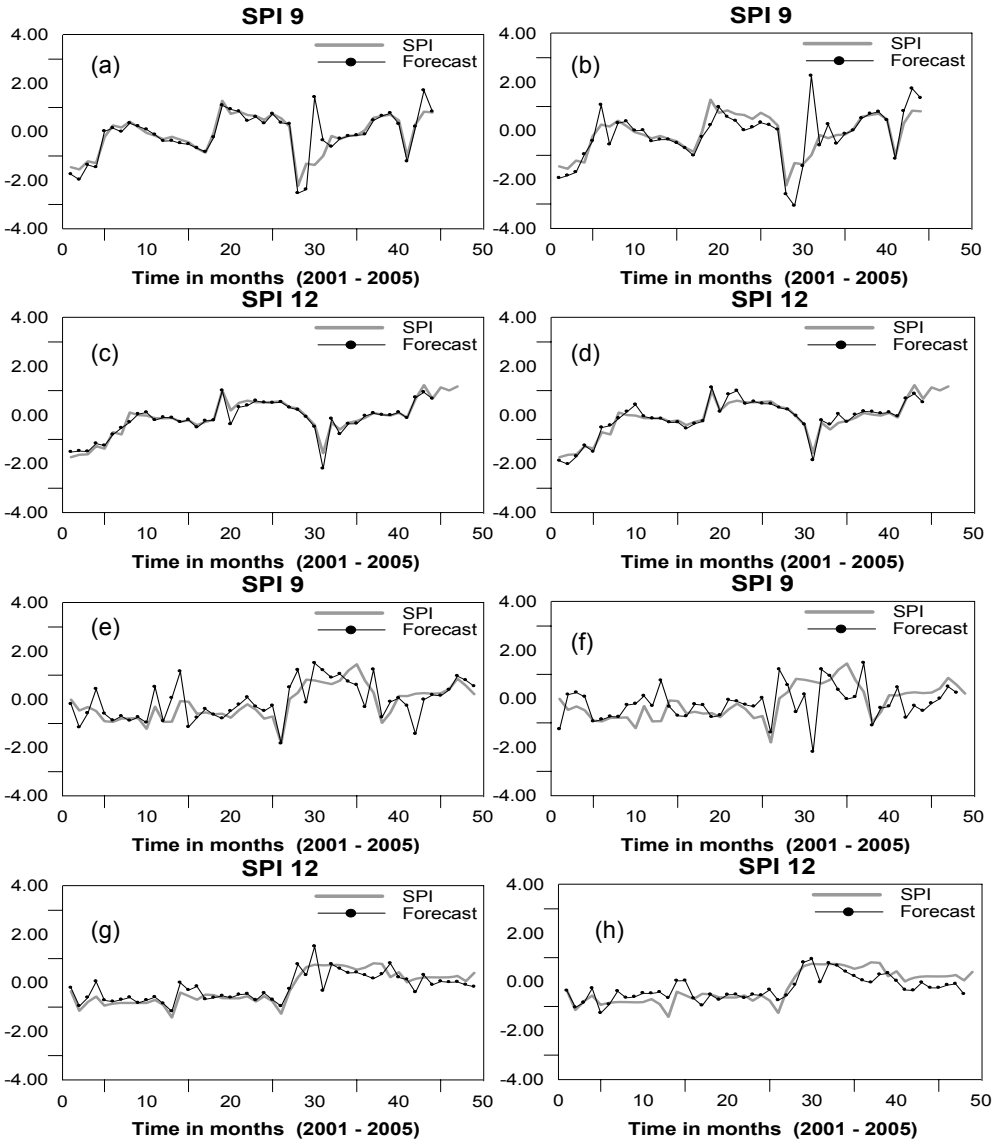

Fig. 7 The SPI-9 and SPI-12 forecast for one month (left) and two months (right): (a)–(d) for Region A, and (e)–(h) for the transition zone.

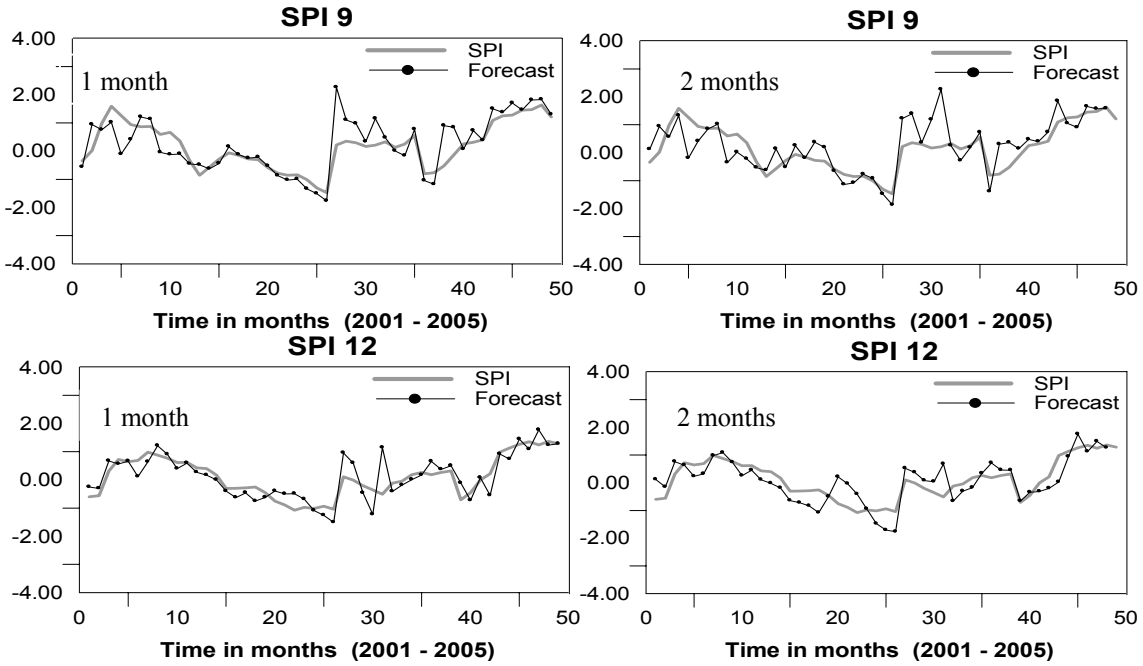

Fig. 8 SPI-9 and SPI-12 forecast for one and two months for Region B.

CONCLUSIONS

Based on the applied methodology and on the results obtained, we can conclude that the proposed ANN proved to be an effective tool in drought forecasting, and that significant influences of the rainfall pattern of the region were not observed in the results. Data from 248 raingauges were analysed and the results of the overall power spectra showed a high annual frequency throughout the basin; however, other frequencies are present with minor significance which represent changes in the rainfall regime. Although, the computed global wavelet power spectra presented this annual frequency, they showed peculiar patterns which could be used to characterize the region. Thus, three sub-regions with homogeneous rainfall patterns were identified as: Region A with frequency pattern A (south part of the basin); Region B with frequency pattern A (north part of the basin); and a transition zone at the central part just between both regions A and B with frequency patterns A and B. Analysing the obtained forecast errors, it is possible to verify that, for 1 month forecast, the indices from 3 to 12 were satisfactory, in which SPI-12 is the most adequate for the forecast. For longer forecasting from three to six months, only SPI-12 presented consistent results and the other indices could be considered inadequate for those horizons, since SPI-12 translates the rainy pattern of the region taking into account longer time scales. However, for the forecast of 3, 4, 5 and 6 months, the results obtained by ANN were not able of represent the drought tendency in the region, even when it was used SPI-12. Bearing in mind that the drought is a phenomenon that reaches great part of the world population, not only in northeastern Brazil, this type of work contributes to significantly understand this phenomenon, which could facilitate the implementation of political actions to fit the real climatic conditions. Further investigations will be able to embody other kinds of ANN, besides other rainfall data and climatological variables, also using GIS and remote sensing techniques, as proposed by Silva *et al.* (2007).

Acknowledgements The authors are grateful to Dr Christopher Torrence of Advanced Study Program at National Center for Atmospheric Research, Colorado, and to the Brazil National Water Agency (ANA), for providing the wavelet analysis computer program and the hydrological data, respectively. This research received financial support from the National Council for Technological and Scientific Development (CNPq - Brazil).

REFERENCES

Giddings, L., Soto, M., Rutherford, B. M. & Maarouf, A. (2005) Standardized Precipitation Index zones for México. *Atmósfera* **18**(1), 35–56.

Guttman, N. B. (1999) Accepting the Standardized Precipitation Index: a calculation algorithm. *J. Am. Water Resour. Assoc.* **35**, 311–322.

Kim, T. W., Valdés, J. B., Nijssen, B. & Roncayolo, D. (2006) Quantification of linkages between large-scale climatic patterns and precipitation in the Colorado River Basin. *J. Hydrol.* **321**, 173–186.

Lloyd-Hughes, B. & Saunders, M. A. (2002) A drought climatology for Europe. *Int. J. Climatology* **22**, 1571–1592.

McKkee, T. B., Doesken, N. J. & Kliest, J. (1993) The relationship of drought frequency and duration to time scales. In: *Proc. Eighth Conf. on Appl. Climatol.* (Anaheim, California, USA, 17–22 January, 1993), 179–184. Am. Met. Soc., Boston, Massachusetts, USA.

McKee, T. B., Doesken, N. J. & Kliest, J. (1995) Drought monitoring with multiple time scales. In: *Proc. of the Ninth Conf. on Appl. Climatol*, (Dallas, Texas, USA), 233–236. Am. Met. Soc., Boston, Massachusetts, USA.

Mishra, A. K. & Desai, V. R. (2006) Drought forecasting using feed-forward recursive neural network. *Ecol. Modell.* **198**, 127–138.

NDMC (National Drought Mitigation Center) (2007) http://www.drought.unl.edu/ (accessed 5 July 2009).

Santos, C. A. G., Srinivasan, V. S., Suzuki, K. & Watanabe, M. (2003) Application of an optimization technique to a physically based erosion model. *Hydrol. Processes* **47**, 989–1003, doi:10.1002/hyp.1176.

Silva, R. M., Santos, C. A. G. & Silva, L. P. (2007) Evaluation of soil loss in Guaraíra basin by GIS and remote sensing based model. *J. Urban Environ. Engng* **1**(2), 44–52, doi:10.4090/juee.2007.v1n2.044052.

Szalai, S., Szinell, Cs. & Zoboki, J. (2000) Drought monitoring in Hungary. *Proc. Expert Group Meeting on Early Warning Systems for Drought Preparedness and Drought Management*, 161–176. World Meteorological Organization, Geneva, Switzerland.

Wu, H., Hayes, M. J., Wilhite, D. A. & Svoboda, M. D. (2005) The effect of the length of record on the Standardized Precipitation Index calculation. *Int. J. Climatol.* **25**, 505–520.