

PROBABILISTIC STRUCTURE OF AN ANNUAL EXTREME RAINFALL SERIES OF A COASTAL AREA OF THE STATE OF SÃO PAULO, BRAZIL

GABRIEL C. BLAIN¹, MARCELO B. P. DE CAMARGO²

ABSTRACT: This study aimed to describe the probabilistic structure of the annual series of extreme daily rainfall (Preabs), available from the weather station of Ubatuba, State of São Paulo, Brazil (1935-2009), by using the general distribution of extreme value (GEV). The autocorrelation function, the Mann-Kendall test, and the wavelet analysis were used in order to evaluate the presence of serial correlations, trends, and periodical components. Considering the results obtained using these three statistical methods, it was possible to assume the hypothesis that this temporal series is free from persistence, trends, and periodicals components. Based on quantitative and qualitative adhesion tests, it was found that the GEV may be used in order to quantify the probabilities of the Preabs data. The best results of GEV were obtained when the parameters of this function were estimated using the method of maximum likelihood. The method of L-moments has also shown satisfactory results.

KEYWORDS: extreme values, adhesion tests, time series.

ESTRUTURA PROBABILÍSTICA DE UMA SÉRIE ANUAL DE PRECIPITAÇÃO PLUVIAL EXTREMA DE UMA REGIÃO DO LITORAL DO ESTADO DE SÃO PAULO

RESUMO: O objetivo do trabalho foi descrever a estrutura probabilística da série anual de precipitação pluvial extrema diária (Preabs), da estação meteorológica de Ubatuba, Estado de São Paulo, Brasil (1935-2009), empregando a distribuição geral dos valores extremos (GEV). A fim de avaliar a presença de correlações seriais, tendências e periodicidades, empregou-se a função auto-correlação, o teste de Mann-Kendall e a análise de ondaletas. Considerando os resultados obtidos com estes três métodos estatísticos, foi possível aceitar as hipóteses de que essa série temporal é livre de persistência, tendências e componentes periódicas. Com base em testes de aderência qualitativos e quantitativos, verificou-se que a GEV pode ser usada para quantificar as probabilidades associadas aos dados de Preabs. A GEV apresentou melhor desempenho quando os parâmetros desta função foram estimados pelo método da máxima verossimilhança. O método dos momentos-L também apresentou desempenho satisfatório.

PALAVRAS-CHAVE: valores extremos, testes de aderência, séries temporais.

INTRODUCTION

Extreme rainfall events are one of the major concerns of the human society due to its potential for causing material damages and human life losses. As pointed out by VICENTE & NUNES (2004) floods (in general triggered by rainfall) may cause, among others hazards, degradation of ecosystems, soil erosion, extensive damage to properties, destruction of crops and (even) trigger slope failures.

As indicated by EL ADLOUNI et al. (2007), the extreme value theory allow us to infer that the probability of occurrence associated with maximum daily rainfall amounts can be estimated by one of the three extreme value distributions (type I - Gumbel; type II - Fréchet and; type III - Weibull). Following WILKS (2006), EL ADLOUNI et al. (2007) and, NADARAJAH & CHOI (2007), these three types can be generalized as a three parameter function called General Extreme

¹ Instituto Agronômico, Centro de Pesquisa e Desenvolvimento de Ecofisiologia e Biofísica, Campinas -SP, Brasil.

² Instituto Agronômico, Centro de Pesquisa e Desenvolvimento de Ecofisiologia e Biofísica, Campinas - SP, Brasil. Bolsista Produtividade CNPq.

Recebido pelo Conselho Editorial em: 24-11-2010

Aprovado pelo Conselho Editorial em: 9-1-2012

Value distribution (GEV). For instance, NADARAJAH & CHOI (2007), PUJOL et al. (2007) and, BLAIN (2010 a), respectively, applied the GEV in order to evaluate the probability associated with maximum precipitation series available from weather stations across the French Mediterranean region, South Korea and, the State of São Paulo, Brazil.

As pointed out by EL ADLOUNI et al. (2007), several methods, such as the maximum likelihood (ML), the L-moments (L) and, the LH-moments (LH) have been proposed for estimating the parameters of the GEV. By setting the shape parameter equal to zero, one is choosing the Gumbel distribution. Positive values of this parameter characterize the Weibull distribution. Negative values of the shape parameter characterize the Fréchet distribution.

Based on several goodness-of-fit tests, BLAIN (2010 a) indicated that the GEV fitted the annual extreme daily rainfall data (Preabs) of the weather station of Ubatuba (1948-2007) better than others distributions, such as normal, lognormal, and 2-parameter gamma. However, there are at least two considerations that have to be discussed in the analysis of the conclusions proposed by these two authors. The first one is related to the length of the rainfall records available for this weather station which begins in 1935 (13 years earlier than the beginning of the time series analyzed by BLAIN, 2010a). The second one is related to the fact that BLAIN (2010a) assumed *a priori* that the ML is the best method for estimating the parameters of the GEV. No consideration about the L and the LH was carried out.

However, as can be verified in the scientific literature, there is still no consensus about which one is the best method for estimating the parameters of the GEV. For instance, according to VALVERDE et al. (2004), the parameters of some distributions, such as the GEV, should be estimated using the L-moments derived from the moments weighted by probability. In addition, as described by QUEIROZ & CHAUDHRY (2006) the LH can be used for evaluating the upper tail of the GEV distributions. On the other hand, for SANSIGOLO (2008) the ML is known as the best estimation parameter method of a probability density function. According to WILKS (2006), for moderate and large sample sizes, the results of the L and the ML are similar.

Thus, the aim of the work was to describe the probabilistic structure of the Preabs time series, available from the weather station of Ubatuba, State of Sao Paulo, Brazil (1935-2009), by using the GEV distribution. The parameters of this distribution were estimated from three methods (ML, L, and LH). The effect of adopting these three different parameters estimation algorithm on the capability of the GEV in describing the empirical distribution was also evaluated.

MATERIALS AND METHODS

Annual extreme daily rainfall data were used from the weather station of Ubatuba, State of São Paulo Brazil, between 1935 and 2009. The weather station is situated at the coastal area where there is no dry season. The annual average rainfall is (approximately) 2650 mm. According to BLAIN (2009), the rainfall monthly series observed in this location are 2-parameter gamma distributed. The parameters of shape and scale of those 12 functions vary from 2 and 30, respectively, (month of July) to 6 and 50 (month of January). Since this time series can be considerate one of the oldest continuous data record available from the coastal area of the State, the evaluation of the presence of trends and periodical components within this data sample may aggregate substantive information to the climate literature of the State of São Paulo. In addition, the knowledge of the rainfall distribution in time-space domain plays an important roll in activities related to agriculture, civil engineering, transports and, tourism (ZANETTI et al., 2005).

As pointed out by MAIA et al. (2007), fitting a cumulative distribution function (cdf) is only appropriate if the time series is not significantly auto-correlated. A cdf summary will result in loss of some information if the time series is moderately to strongly auto-correlated (MAIA et al., 2007). Thus, the auto-correlation function (ACF) was used in order to verify if the data sample can be considerate as generated from a white noise process. The coefficients of the ACF were estimated following WILKS (2006) from lags 1 to 12 (years). It is worth it to mention that the presence of trends and/or periodicals components may affect the probabilistic structure of the data sample.

Thus, two statistical methods were applied in order to evaluate the presence of these components within the time series.

The Mann-Kendall (MK) test (MANN, 1945; KENDALL & STUART, 1967) is largely used for evaluating the presence of trends in meteorological time series (BLAIN et al., 2009; BLAIN, 2010 b; PUJOL et al., 2007 and, SANSIGOLO & KAYANO, 2010). The null hypothesis (H0) associated with this test assumes that the sample is free from trends (the absence of significant serial correlation is also assumed). The H0 is usually rejected if the p-value is less than or equal to 0.05. The MK was used as a methodology for trend estimation within the Preabs series.

Following TORRENCE & COMPO (1998) the wavelet analysis was used in order to decomposing the Preabs time series into time-frequency space. Thus, this form of spectral analysis has allowed us to (i) observe the variance peaks in the frequency domain and to (ii) verify how those peaks vary in time. Detailed explanation of the wavelet technique can be found in TORRENCE & COMPO (1998). Following BLAIN (2010a), the wavelet function (mother wavelet) used in the present study was the Morlet. The wavelet analysis (including the statistical significance testing) was estimated from the computational procedure described by TORRENCE & COMPO (1998) and available at <http://paos.colorado.edu/research/wavelets> (accessed at November 30, 2010).

The GEV can be described as:

$$f(x) = \frac{1}{\beta} \left[1 + \frac{k(x-\zeta)^{-1-\frac{1}{k}}}{\beta} \right] \exp \left\{ - \left[1 + \frac{k(x-\zeta)^{-1-\frac{1}{k}}}{\beta} \right]^{\frac{1}{k}} \right\} \quad \text{only if } 1 + \frac{k(x-\zeta)^{-1-\frac{1}{k}}}{\beta} > 0 \quad (1)$$

Where:

- ζ - location parameter;
- β - scale parameter, and
- k - shape parameter.

The corresponding quantile (p) function can be estimated by:

$$F^{-1}(p) = \zeta + \frac{\beta}{k} \{ [-\ln(p)]^{-k} - 1 \} \quad (2)$$

For NADARAJAH & CHOI (2007), the GEV has all the flexibility of its three particular types. The parameters of equation 1 were estimated using the methods of ML, L, and LH. As described by QUEIROZ & CHAUDHRY (2006), the LH is based on linear combinations of higher order probability weight moments. When the order of LH is $\eta=0$, its value becomes equivalent to L (it were considered $\eta=0, 1, 2, 3$ and 4).

The chi-square test (χ^2) and the Kolmogorov-Smirnov test (KS) were used to verify if the Preabs series were drawn from a GEV distribution. As pointed out by WILKS (2006) the χ^2 test actually operates more naturally for discrete random variable since to calculate it, the range of the data must be divided into discrete classes. For continuous distributions because the KS test compares the empirical and the theoretical cumulative functions it is frequently more powerful than the χ^2 test. The H0 associated with these both tests assumes that the data sample under evaluation was drawn from a hypothesized (GEV) distribution.

However, it is worth it to mention that as discussed by WILKS (2006), STEINSKOG et al. (2007) and, VLCEK & HUTH (2009), if (and only if) the parameters of the theoretical distribution have not been estimated from the same data sample used to evaluate the fit of the parametric distribution, the original algorithm of the Kolmogorov-Smirnov test is applicable. Thus, since the three parameters of the GEV were fitted using all available data, the KS test had to be modified. Hereafter this adapted method will be referred as Kolmogorov-Smirnov/Lilliefors test (KS-L). The statistical simulations required for calculating the KS-L test were based on the procedure called

“nonuniform random number generation by inversion”. It were generated $N_s=100000$ synthetic data samples. More information about the χ^2 and the KS-L can be found in WILKS (2006).

Although the KS-L and the χ^2 are commonly used tests of goodness-of-fit, these both methods are only appropriated for evaluating the central part of the distributions (SANSIGOLO, 2008). Since the aim of the study is to evaluate extreme rainfall amounts, it becomes evident that special focus should be given for the upper tail of the distributions. Furthermore, according to WILKS (2006), although formal tests (such as KS-L and χ^2) may indicate an inadequate fit, they may not inform the researcher as to the specific nature of the problem. In this view, the quantile-quantile plots (QQ), as described by WILKS (2006) were used in order to compare the observed data and the fitted distribution. Thus, the QQ plots have allowed us to verify how and where the parametric representation was not adequate. Finally, although the QQ plots are usually classified as a qualitatively procedure of assessing the goodness-of-fit (WILKS, 2006), the mean absolute error (MAE) and the mean squared error (MSE) were used in order to support the evaluation of the results of these plots. Following WILKS (2006) the MAE is zero if the fit is perfect, and increases as the discrepancies between the empirical and the theoretical quantile become larger. The MSE is similar to the MAE but the squaring function is used rather than the absolute value. Thus the MSE is more sensitive to larger discrepancies than the MAE. The MSE was also expressed as its square root ($RMSE=\sqrt{MSE}$).

RESULTS AND DISCUSSION

The ACF allowed us to accept the hypothesis that the Preabs series is generated from a white noise process, since all the coefficients of this function fell within the white noise limits (Figure 1). Thus, following MAIA et al. (2007), a cdf summary of the Preabs series will result in loss of no information, since no (significant) persistence was observed within the data sample. This lack of serial correlation also allowed us to evaluate the presence of trends in the Preabs series by using the (original) MK algorithm. No adaptation due the presence of persistence had to be adopted in the MK algorithm (for more information see HAMED & RAO, 1998). Since the p-value associated with the MK final value (Figure 1) is far from the critical limit ($p \leq 0.05$) there were no statistical evidences to reject H_0 .

The wavelet analysis shows the absence of significant periodical components within the Preabs time series. For instance, no significant peak (at 5% level; represented by the dashed line) can be observed in Figure 2b (global wavelet power; GWP). Furthermore, the wavelet power spectrum (WPS; Figure 2a) shows concentration of energy only during small periods of time. For the 2-4 year band, there is appreciable power during the beginning of the 1940s (at the border of the “cone of influence”) and between the years of 1968 and 1980 (approximately). Another concentration of wavelet power can only be verified at the beginning of the series between the 8-12 year bands.

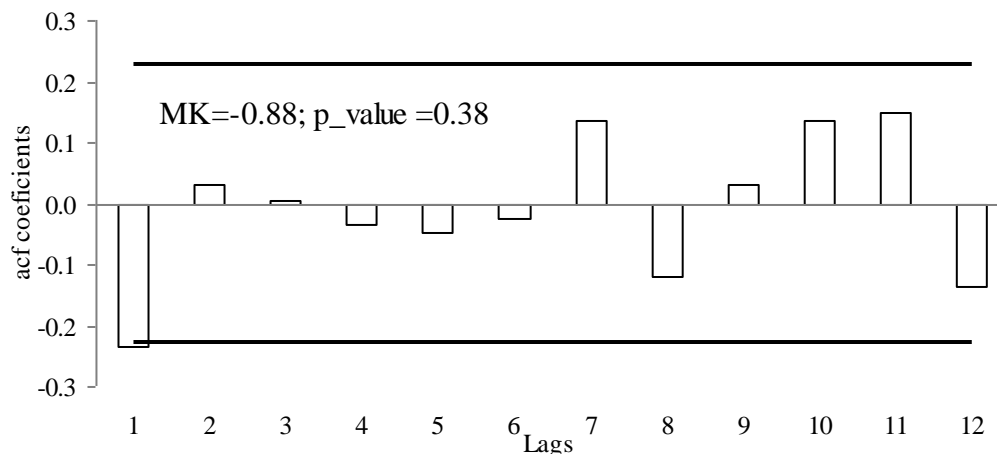


FIGURE 1. Auto-correlation function (acf) applied to the annual extreme rainfall series of the location of Ubatuba (1935-2009), SP. The Mann-Kendall test (MK) is also illustrated.

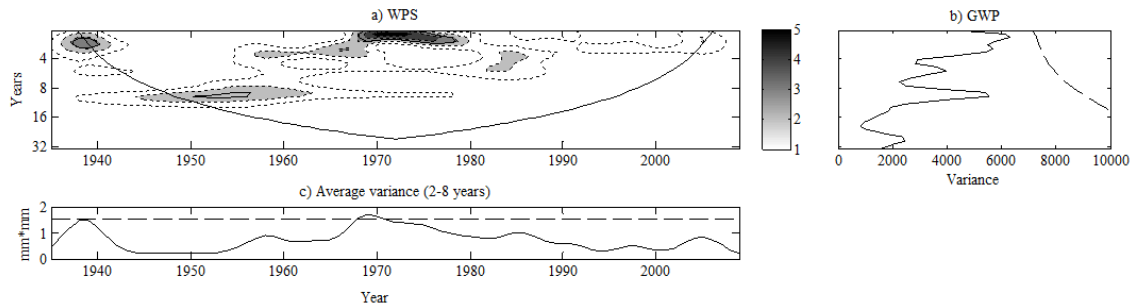


FIGURE 2. (a) Wavelet power spectrum (WPS) of the annual extreme rainfall series of the location of Ubatuba (1935-2009), SP, (normalized by $1/4627\text{mm}^2$); (b) Global Power Spectrum (GWP - in variance units); (c) Scale-average wavelet power over 2-8 year band.

Considering the results depicted in Figures 1 and 2, it becomes reasonable to assume that the data sample under evaluation is a sequence of independent and identically distributed variables. Following CALGARO et al., (2009) the data sample used in this study can be considerate as following a random process. Thus, it also becomes reasonable to evaluate the possibility of using a parametric distribution (stochastic model) in order to assess the probability of occurrence associated with these variables. Since the present study deals with maximum daily rainfall amounts, the general extreme value distribution has become a natural choice. As indicated by the χ^2 test, the GEV can be used to assess the probability of occurrence of the Preabs values only if the parameters of equation 1 are estimated using either the L method ($\eta=0$) or the ML method. The inadequacy of adopting the LH method ($\eta=1$ to 4) resulted in values of the χ^2 test associated with $p < 0.01$ (H_0 could be rejected). These last results can be easily observed in Figure 3.

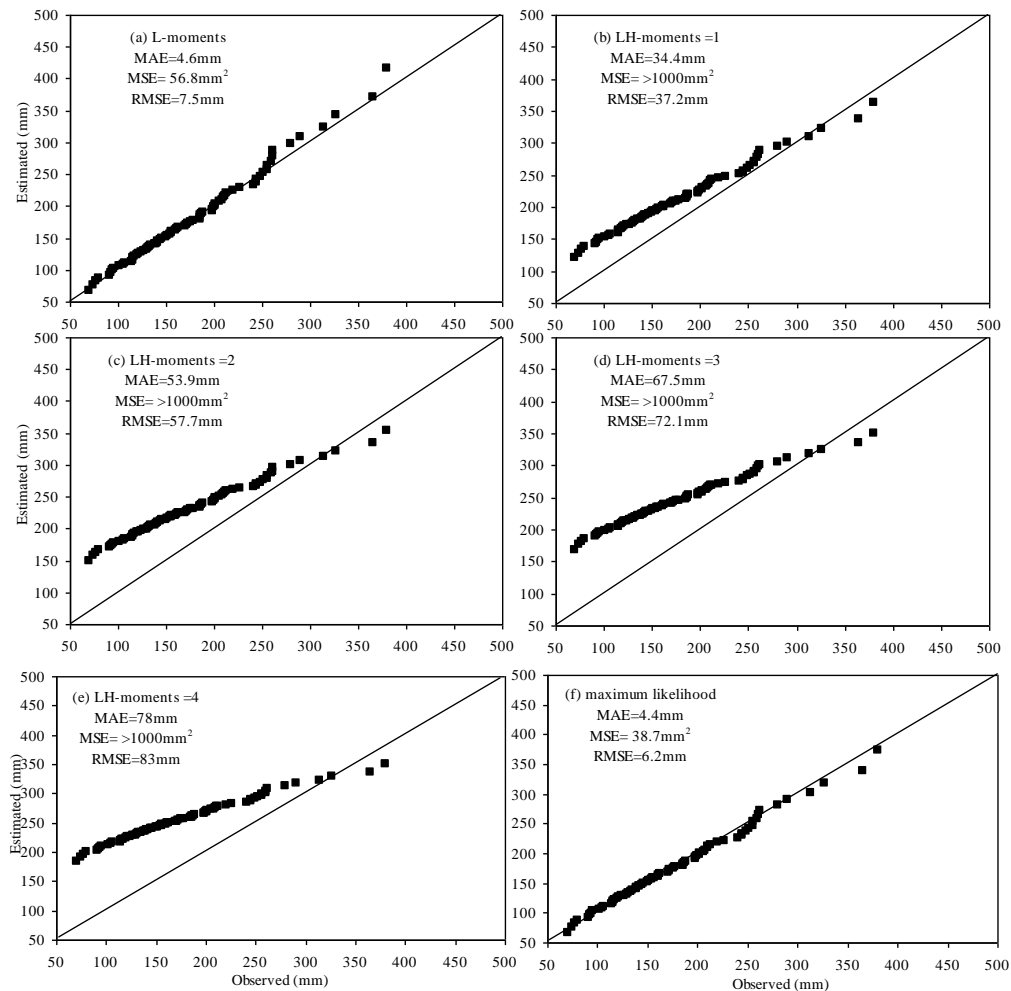


FIGURE 3. Quantile-quantile plots. Observed annual extreme rainfall data of the location of Ubatuba (1935-2009), SP (x axes). Quantile function associated with the general extreme value distribution (y axes). The mean absolute error (MAE), mean square error (MSE) and its square root (RMSE) are also shown.

Figure 3 also suggests that although the L method has presented satisfactory results, the ML can be seen as the best estimation parameter method of the GEV distribution fitted from the Preabs data sample. The lowest values of the scalar accuracy measures are observed in Figure 3f. The same conclusion is obtained when the results of the KS-L are evaluated. As can be seen in Figure 4, the maximum absolute difference between the parametric distribution and the empirical distribution ($D_{max}=0.0496$) is associated with a significance level ($p=0.45$) far from the commonly adopted probability rejection level ($p=0.05$; maximum probability of occurrence of error type I). Adopting the L method has resulted in $D_{max} = 0.0548$ ($p=0.42$). Thus, once again, although the L method has presented a satisfactory result, the ML was a slightly better model for estimating of the parameters of GEV distribution.

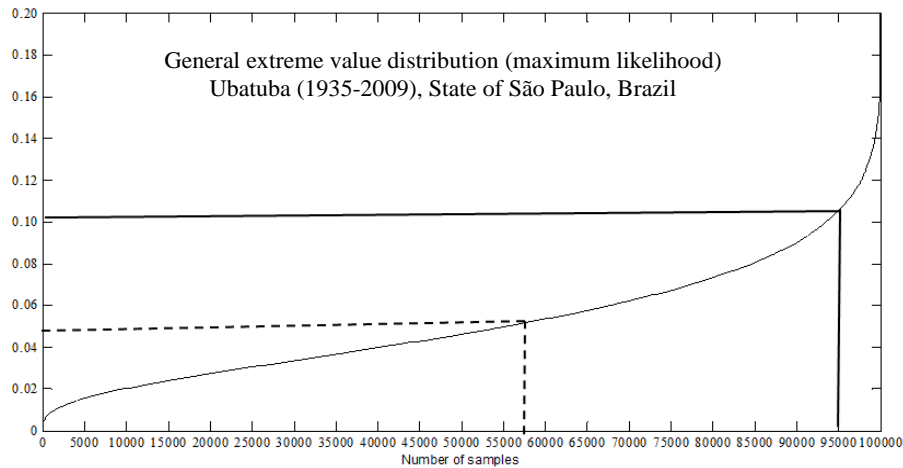


FIGURE 4. Kolmogorov-Smirnov/Lilliefors goodness-of-fit test. The solid line represents the 5% (usually adopted) significance level. The dashed line represents the maximum absolute difference between the parametric distribution and the empirical distribution.

By adopting the ML we have obtained parameters of the equation 1 equal to: $\zeta = 147.464$; $\beta = 54.740$ and, $k = -0.0235$. By adopting the L we have obtained parameters of the equation 1 equal to: $\zeta = 147.467$; $\beta = 56.952$ and, $k = 0.0411$. The results depicted in Figures 1 to 4 have allowed us to use the GEV in calculating the probability of occurrence associated with the Preabs values of the location of Ubatuba. The return period $\{1/[1-f(x)]\}$ corresponding to each Preabs value are also shown.

As can be verified in Figure 5, by using the information presented in this study, one is able to i) estimate the probability of occurrence associated with a chosen Preabs value, ii) estimate the return period associated with this Preabs value and, iii) estimate the value of Preabs associated with a chosen probability level (equation 2). In addition, one is also able to assume that the analyzed data sample is free from significant serial correlations, periodicals components and, trends.

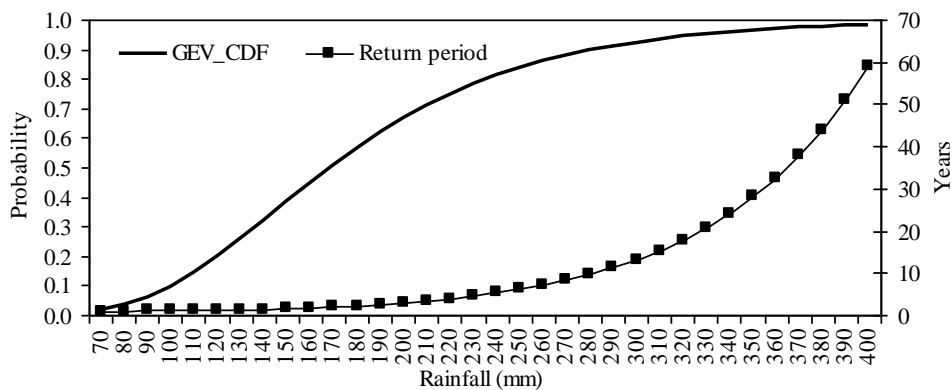


FIGURE 5. Cumulative probability and associated return period of the annual extreme rainfall values - Ubatuba (1935-2009), SP.

CONCLUSIONS

The time series composed from annual extreme daily rainfall data of the weather station of Ubatuba (1935-2009) can be considered free from significant temporal persistence. Neither significant trends nor periodical components were observed within this data sample.

The general extreme value distribution can be used in order to evaluate the probabilistic structure of this time series. The best results of the GEV were obtained when the parameters of this function were estimated using the maximum likelihood method. The method of L-moments has also shown satisfactory results. The LH method ($\eta=1, 2, 3$ and, 4) cannot be recommended.

REFERENCES

- BLAIN, G.C. Considerações estatísticas relativas à oito séries de precipitação pluvial da Secretaria de Agricultura e Abastecimento do Estado de São Paulo. *Revista Brasileira de Meteorologia*, São José dos Campos, v.24, n.1, p.12-23, 2009.
- BLAIN, G.C. *Precipitação pluvial e temperatura do ar no Estado de São Paulo: periodicidades, probabilidades associadas, tendências e variações climáticas*. 2010. 194 f. Tese (Doutorado em Agronomia) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2010a.
- BLAIN, G.C. Séries anuais de temperatura máxima média do ar no Estado de São Paulo: variações e tendências climáticas. *Revista Brasileira de Meteorologia*, São José dos Campos, v.25, n.1, p.114-124, 2010b.
- BLAIN, G.C.; ARAUJO, M.C.; LULU, J. Análises estatísticas das tendências de elevação nas séries anuais de temperatura mínima do ar no Estado de São Paulo. *Bragantia*, Campinas, v.68, n.3, p.807-815, 2009
- CALGARO, M.; ROBAINA, A.D.; PEITER, M.X.; BERNARDON, T. Variação espaço-temporal dos parâmetros para a modelagem estocástica da precipitação pluvial diária no Rio Grande do Sul. *Engenharia Agrícola*, Jaboticabal, v.29, n.2, p.196-206, 2009.
- EL ADLOUNI, S.; OUARDA, T.B.M.J.; ZHANG, X.; ROY, R.; BOBÉE, B. Generalized maximum likelihood estimators for the nonstationary generalized extreme value model. *Water Resources Research*, v.43.W03410. Disponível em: <<http://dx.doi.org>>. DOI:10.1029/2005WR004545, 2007
- HAMED, K.H.; RAO, A.R. A modified Mann-Kendall trend test for auto-correlated data. *Journal of Hydrologic*, Reston, v.204, p.182-196, 1998.
- KENDALL, M.A.; STUART, A. *The advanced theory of statistics*. 2.ed. Londres: Charles Griffin & Company, v.2, 1967. 690 p.
- MAIA, A.H.N.; MEINKE, H.; LENNOX, S.; STONE, R.C. Inferential, non-parametric statistics to assess quality of probabilistic forecast systems. *Monthly Weather Review*, Boston, v.135, p.351-362, 2007.
- MANN, H.B. Non-parametric tests against trend. *Econometrica*, Chicago, v.13, p.245-259, 1945.
- NADARAJAH, S.; CHOI, D. Maximum daily rainfall in South Korea. *Journal of Earth System Science*, v.116, n.4, p.311-320, 2007.
- PUJOL, N.; NEPPEL, L.; SABATIER, R. Regional tests for trend detection in maximum precipitation series in the French Mediterranean region. *Hydrological Sciences Journal*, Oxford, v.52, n.5, p.956-973, 2007.
- QUEIROZ, M.M.F.; CHAUDHRY, F.H. Análise de eventos hidrológicos extremos, usando-se a distribuição GEV e momentos LH. *Revista Brasileira de Engenharia Agrícola e Ambiental*, Campina Grande, v.10, n.2, 2006.

- SANSIGOLO, C. A. Distribuições de extremos de precipitação diária, temperatura máxima e mínima e velocidade do vento em Piracicaba, SP (1917-2006). *Revista Brasileira de Meteorologia*, São José dos Campos, v.23, n.3, p.341-346, 2008.
- SANSIGOLO, C.A.; KAYANO, M.T. Trends of seasonal maximum and minimum temperatures and precipitation in Southern Brazil for the 1913-2006 period, *Theoretical and Applied Climatology*, Wien, v.101, p.209-216, 2010.
- STEINSKOG, D.J.; TJØSTHEIM, D.B.; KVAMSTØ, N.G. A cautionary note on the use of the Kolmogorov–Smirnov test for normality. *Monthly Weather Review*, Boston, v.135, n.3, p.1151-1157, 2007.
- TORRENCE, C.; COMPO, G.P. Practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, Boston, v.79, p.61-78, 1998.
- VALVERDE, A.E.L.; LEITE, H.G.; SILVA, D.D.; PRUSKI, F.F. Momentos-L: teoria e aplicações em hidrologia. *Revista Árvore*, Viçosa-MG, v.28, n.6, p.927-933, 2004.
- VICENTE, A.K.; NUNES, L.H. Extreme precipitation events in Campinas, Brazil. *Terrae*, Campinas, v.1, n.1 p.60-62, 2004.
- VLČEK, O.; HUTH R. Is daily precipitation Gamma-distributed? Adverse effects of an incorrect application of the Kolmogorov-Smirnov test. *Atmospheric Research*, Amsterdam, v. 93, n.4, p.759-766, 2009.
- WILKS, D.S. *Statistical methods in the atmospheric sciences*. 2nd.ed. San Diego: Academic Press, 2006. 629 p.
- ZANETTI, S.S.; PRUSKI, F.F.; MOREIRA, M.C.; SEDIYAMA, G.C.; SILVA, D.D. Programa computacional para geração de séries sintéticas de precipitação. *Engenharia Agrícola*, Jaboticabal, v.25, n.1, p.96-104, 2005.