Severe Thunderstorms
Extreme values analysis

Nicolas Rime Matthieu Baud
n.rime@hotmail.com matthieu.baud@epfl.ch

January 6, 2019

Summary: Violent thunderstorms, including tornadoes can cause a huge amount of damages to populated areas, hence the repair costs can be highly expensive for businesses and insurance companies. The aim of this project is to study the extremal values of the combination of the Convective Available Potential Energy and the Storm Relative Helicity, in order to assess the risk of such unfortunate events. The contribution of the authors is 80% for Nicolas Rime and 20% for Matthieu Baud, decided by agreement between the two authors.

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Contents

1 Introduction

1.1 Purposes

The purpose of this study is to analyze the extremal properties of the variable $PROD = \sqrt{CAPE} \times SRH$, where CAPE means Convective Available Potential Energy and SRH means Storm Relative Helicity. High values of CAPE and SRH have proven to be important for the risk analysis of severe thunderstorms. Furthermore, an extremal dependence structure analysis of these two variables will be applied.

1.2 Data

Convective Available Potential Energy (CAPE) is the amount of energy that an air surface would have if it would be lifted vertically through the atmosphere. The severe weather potential at a given area at a given time can be determined by this quantity [1]. The CAPE data set contains a large number of zero values, meaning that the potential energy is equal to zero in this area and therefore the weather is calm during these measurements. On the other hand, extreme values of CAPE can result in violent thunderstorm developments.

Storm Relative Helicity (SRH) is a measure of the potential for cyclonic updraft rotation in right-moving supercells. A supercell is a thunderstorm characterized by the presence of a mesocyclone: a deep, persistently rotating updraft. Greater is the value of SRH for a supercell, greater is the risk of tornado but it does not exist a theoretical threshold to determine at which value the risk is significant [2].

These two information are very important for the measure of the risk that a thunderstorm happens over an area. Each data set contains 108'040 values corresponding to eight measurements per days (every three hours) during 37 years from 1979 to 2015 (February 29 of each leap year are ignored). The values correspond to measures made in a grid cell whose South-West corner has coordinates 35° latitude, -101° longitude. The grid cell are 1° latitude, 1° longitude.

We also have access to the NINO 3.4 index data set, which is a good indicator of the El Nino-Southern Oscillation (ENSO). ENSO is an irregularly periodic variation in winds and sea surface temperatures over the tropical eastern Pacific Ocean, affecting the climate of much of the tropics and subtropics. The Nino 3.4 is an index to define the warmest and the coldest phases of the ENSO: El Nino and La Nina.

1.3 Preprocess

After applying the formula to compute the $PROD$ variable, the PROD data set contains 108'040 values. To work on extremal models, the data points are pre-processed to analyze only the maximum month values as shown in Figure 1. Finally the resulting data set contains 444 extremal month values. As maximum values changes over the months, the maximum per each month are considered for analysis, giving 37 maximums per month. The PROD points tends to be specifically high around the 4^{th} month, April.

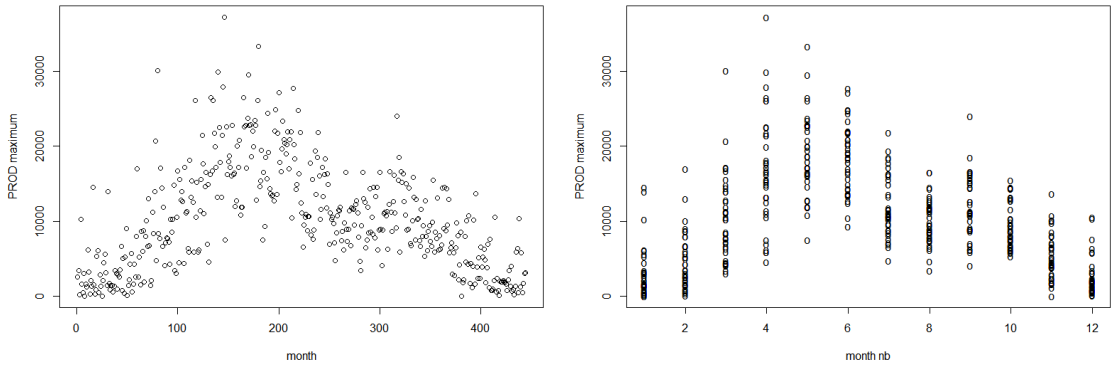

Figure 1: Maximum per month of PROD, for all months (left) and each month separated (right)

The ENSO data set is pre-processed in the same way. The ENSO values per month during 37 years are shown on figure 2.

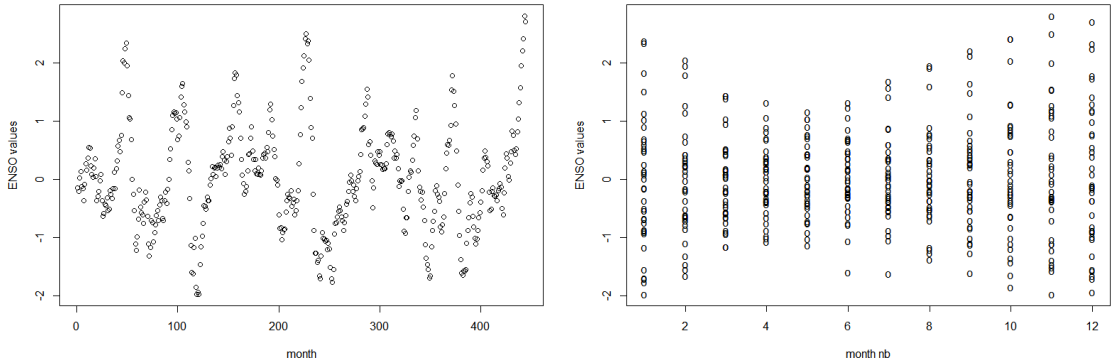

Figure 2: Values of ENSO per month, for all months (left) and each month separated (right)

2 Fit of the Generalized Extreme Value distribution (GEV)

In order to fit the maxima per month, three different approaches will be considered: maximum likelihood, Bayesian techniques and r-order statistics. The GEV distribution function is

$$G(x) = \exp(-[1 + \xi(x - \eta)/\tau]_+^{-1/\xi}) \tag{1}$$

where η, τ and ξ are respectively the location, scale and shape parameters that must be estimated and for a real, $a_+ = \max(a, 0)$.

The first method consists of finding the parameters which maximize the log-likelihood function, meaning the logarithm of the GEV distribution function. The second uses Markov chains Monte Carlo (abbreviated MCMC), more precisely the Metropolis-Hastings algorithm. The last method fits the r largest maximum of each block considered.

		Jan	Feb	Mar	Apr	May	Jun
η $[10^3]$	max-lh	-	$2.99_{0.52}$	$7.93_{0.76}$	$13.4_{1.16}$	$16.4_{0.85}$	$16.4_{0.83}$
$[m^3\ s^{-3}]$	MCMC	$1.12_{0.24}$	$2.71_{0.48}$	$7.73_{0.77}$	$13.5_{1.17}$	$16.5_{0.92}$	$16.7_{0.83}$
	$r = 2$	$1.53_{0.22}$	$3.33_{0.40}$	$9.68_{0.66}$	$14.4_{0.93}$	$18.9_{0.72}$	$16.8_{0.65}$
τ $[10^3]$	max-lh	-	$2.89_{0.58}$	$4.35_{0.68}$	$5.79_{0.53}$	$4.37_{0.45}$	$3.62_{0.37}$
$[m^3\ s^{-3}]$	MCMC	$1.37_{0.27}$	$2.49_{0.44}$	$4.15_{0.64}$	$6.85_{0.91}$	$5.36_{0.72}$	$4.83_{0.66}$
	$r = 2$	$1.43_{0.22}$	$2.38_{0.35}$	$4.01_{0.41}$	$6.54_{0.57}$	$4.61_{0.37}$	$4.09_{0.38}$
ξ $[10^{-1}]$	max-lh	-	$1.64_{2.5}$	$0.52_{1.61}$	$-0.83_{1.06}$	$-1.28_{0.87}$	$-1.70_{1.12}$
	MCMC	$5.80_{2.00}$	$2.73_{2.03}$	$1.30_{1.58}$	$-0.7_{1.30}$	$-1.50_{1.00}$	$-2.73_{1.26}$
	$r = 2$	$6.09_{1.66}$	$3.90_{1.61}$	$0.48_{1.06}$	$-0.92_{0.99}$	$-2.04_{0.67}$	$-2.08_{1.07}$

		Jul	Aug	Sep	Oct	Nov	Dec
η $[10^3]$	max-lh	$10.2_{0.63}$	$8.74_{0.5}$	$9.94_{0.64}$	$7.99_{0.46}$	$3.67_{0.47}$	-
$[m^3\ s^{-3}]$	MCMC	$10.2_{0.67}$	$8.98_{0.53}$	$9.98_{0.71}$	$7.79_{0.46}$	$3.64_{0.49}$	$1.20_{0.20}$
	$r = 2$	$10.9_{0.51}$	$10.3_{0.41}$	$11.8_{0.53}$	9.02_{042}	$4.71_{0.40}$	$1.67_{0.20}$
τ $[10^3]$	max-lh	$3.18_{0.39}$	$2.43_{0.23}$	$3.31_{0.37}$	$2.28_{0.40}$	$2.45_{0.35}$	-
$[m^3\ s^{-3}]$	MCMC	$3.63_{0.51}$	$3.20_{0.46}$	$3.91_{0.53}$	$2.24_{0.42}$	$2.62_{0.38}$	$1.12_{0.23}$
	$r = 2$	$3.28_{0.30}$	$2.57_{0.19}$	$3.36_{0.28}$	$2.43_{0.23}$	$2.42_{0.25}$	$1.29_{0.19}$
ξ $[10^{-1}]$	max-lh	$-0.55_{1.26}$	$-1.62_{0.90}$	$-0.82_{1.03}$	$0.83_{2.3}$	$0.01_{1.49}$	-
	MCMC	$-0.64_{1.35}$	$-2.43_{1.26}$	$-0.88_{1.16}$	$1.58_{2.18}$	$0.33_{1.45}$	$5.03_{1.76}$
	$r = 2$	$-0.32_{1.15}$	$-2.86_{0.79}$	$-1.27_{0.81}$	$-0.56_{1.29}$	$0.32_{1.14}$	$4.42_{1.31}$

Table 1: Table of parameters of the GEV per month, with the standard deviation in subscript.

2.1 Maximum likelihood

The first approach use likelihood inference to estimate the parameters of the GEV. Under regularity conditions on data, the maximum likelihood estimator $\hat{\theta}$ satisfies:

$$\hat{\theta} \sim N_p\left(\theta^0, J(\hat{\theta})^{-1}\right),$$

where $J(\theta)$ is the observed information matrix

$$J(\theta) = -\mathbb{E}\left[\frac{\partial^2 l}{\partial \theta^2}\right] \tag{2}$$

and the parameter vector θ is divided into interest and nuisance parameters. For the GEV distribution the log-likelihood function is:

$$l(\eta, \tau, \xi) = \sum_{i=1}^{n}\left[-\log(\tau) - (1+1/\xi)\log\left\{1+\xi\left(\frac{y_i-\eta}{\tau}\right)\right\}_+ - \left\{1+\xi\left(\frac{y_i-\eta}{\tau}\right)\right\}_+^{-1/\xi}\right] \tag{3}$$

where the y_i are the data points. The fitted parameters and standard deviations are presented in table 1. Concerning the months of January and December, the information matrix is singular and hence no estimates are given. The standard deviations given in table 1 can give a symmetric confidence interval, $\theta \pm 1.96 \cdot \sigma_\theta$ for $\theta = \{\eta, \tau, \xi\}$ and σ_θ is the standard deviation of the parameter. In order to have a more accurate and asymmetric confidence interval, the profile log-likelihood should be considered. The diagnostic plots and the profile plots for the month of March is shown

in figure 3. The diagnostic plots for the other months are not shown but all quantile plots indicate that the model fits well. The profile plots for the other months are shown in annex, to show the explicit asymmetric 95% confidence interval.

Diagnostic plots suggest that the model is adequate and it indicates quite uncertainties about the upper tail. Depending of the month this uncertainty varies a lot (greater uncertainty for February for instance and low uncertainty for April). For March, the profile likelihood for η is nearly symmetric so the likelihood-based confidence interval and the normal confidence interval are similar. It is not the case for τ and ξ where the profile likelihood plot is asymmetric.

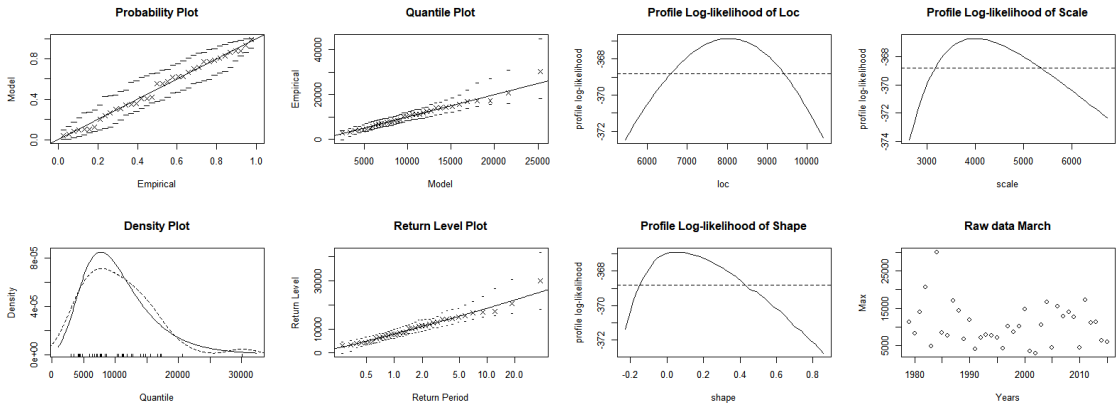

Figure 3: Diagnostic plots, profile log-likelihood and data for March

2.2 Markov chain Monte Carlo and Metropolis-Hastings

The second inference approach uses stochastic computation using Bayesian techniques. The Metropolis-Hastings algorithm generates a Markov Chain $\{\theta^t\}$ with a posterior density π:

1. Set $t = 0$ and choose initial state $\theta^{(0)}$.
2. Simulate θ^* with a predefined density $q(.|\theta^{(t)})$.
3. Calculate acceptance probability:

$$\alpha = \alpha(\theta^{(t)}, \theta^*) = \min\left\{1, \frac{\pi(\theta^*)q(\theta^{(t)}|\theta^*)}{\pi(\theta^{(t)})q(\theta^*|\theta^{(t)})}\right\}$$

4. Define: $\theta^{(t+1)} = \theta^*$ with probability α and $\theta^{(t+1)} = \theta^{(t)}$ with probability $1 - \alpha$.
5. Set $t = t + 1$ and go to 2.

The results of the Markov chains Monte Carlo are in table 1. For the twelve simulations, the same initial values of the estimators are taken, as well as the same uninformative prior, which is simply a Gaussian vector with a large diagonal covariance matrix. The proposal distribution is also Gaussian, however its covariance matrix is adapted to each month in order to have an acceptance rate between 20% and 40 % for each parameters. For 5000 points generated, the 500 first are removed (the initial burn-in) and the auto-correlation function suggest a thinning of the chain by a factor 9. The same thinning is taken for all months.

2.3 r - largest order statistics

Finding the optimal r can be done by analyzing residual plots. A large r takes more data points into account but can gives biased results. After a few tries, $r = 2$ seems reasonable. The results are displayed on table 1.

2.4 Discussion of the fitted parameters

The most important parameter to consider is the shape parameter ξ as its sign determines the type of the distribution. If the shape is positive, zero or negative, the distribution is respectively a Frechet, Gumbel or reverse Weibull distribution. As table 1 shows, the months between April and September have a reverse Weibull distribution and the others months follow a Frechet distribution. However, it is likely that certain months like March, April, July, September, October and November follow a Gumbel distribution as their shape parameter is not so distant from zero and furthermore zero is their 95% confidence interval. The reverse Weibull distribution has a maximum point, meaning that following this model, PROD has a maximum value for the months between April and September given by $\eta - \tau/\xi$. The location parameter η increases over the first six months of the year and then decreases over the last six months. In general, the method of maximum likelihood and MCMC give similar results. MCMC allows to compute parameters for January and December, when the maximum likelihood method failed due to a singular information matrix. Concerning the r-order statistics, the values for $r = 2$ seem to be appropriate.

3 Time, ENSO dependence

The location parameter of the GEV distribution is maybe dependent on time or ENSO. The null hypothesis can be written "*PROD is independent of time and ENSO*". The new hypotheses are *i*) "*PROD is dependent of time*" and *ii*) "*PROD is dependent of ENSO*". Rejection of the null hypothesis can be measured with a ratio likelihood test: LRT = deviance(simple model) - deviance(complex model). A smaller deviance signify a better fit. Asymptotically, this test ratio statistics is distributed as a chi-squared random variable with one degree of freedom. The p-value is equal to $\mathbb{P}(X \geq LRT)$ where X follow a chi-squared distribution with one degree of freedom. For every month, the p-value is high, so it is likely that we can reduce the time or ENSO dependent model to a simpler model, meaning that it is more likely that the null hypothesis is true. Exception is June and August for the dependence with ENSO which give p-values of 0.047 and 0.040 respectively. A second test of the extremal dependance of PROD and ENSO is to compute the correlation of each month. The maximum correlation is 0.237 for August and the minimum is -0.266 for May. Hence PROD and ENSO are weakly correlated for each month.

4 Temporal clustering and return levels

In the last section, the extreme values were considered by blocks. Namely for each month, only one maximum per year was taken into account, giving 37 points. The approach of clustering allows to

consider extreme values over a threshold, giving more points for analysis. Two different methods are used, fit the General Pareto distribution (GPD) and the point process. The GPD is defined by:

$$H(x) = 1 - (1 + \xi x/\sigma)_+^{-1/\xi}, \quad \xi \neq 0 \tag{4}$$

where $\sigma = \tau + \xi(u - \eta) > 0$, u being the threshold. For each month a different threshold is used, reported on table 2. The threshold was chosen after analyzing the mean residual life plot and parameters stability plots, shown on figure 4 for November. Finding an appropriate threshold is important to ensure a good trade-off between bias and variance. Indeed, a low threshold takes more points into account but the estimated shape and scale have more bias. On the other hand, the variance of the estimated parameters can be large if a high threshold is chosen.

		Jan	Feb	Mar	Apr	May	Jun
threshold [10^3]		1.2	4.0	9.0	12	7.5	10
θ		0.5	0.5	0.6	0.7	0.6	0.7
rl 50 [10^4] [$m^3\ s^{-3}$]	pp	1.55	1.42	2.52	3.40	3.22	-
	gpd	1.35	1.42	2.52	3.45	3.33	3.00
	MCMC	1.3	1.4	2.5	3.4	3.1	2.7
rl 100 [10^4] [$m^3\ s^{-3}$]	pp	2.02	1.59	2.80	3.64	3.41	-
	gpd	1.61	1.59	2.79	3.73	3.56	3.17
	MCMC	1.9	1.6	2.8	3.7	3.3	2.8
		Jul	Aug	Sep	Oct	Nov	Dec
threshold [10^3]		12	8	10	8	5	2
θ		0.9	0.8	0.9	0.7	0.6	0.5
rl 50 [10^4] [$m^3\ s^{-3}$]	pp	-	-	2.18	1.53	1.28	-
	gpd	2.08	1.71	2.19	1.54	1.29	1.07
	MCMC	2.1	1.7	2.1	1.7	1.2	1.1
rl 100 [10^4] [$m^3\ s^{-3}$]	pp	-	-	2.29	1.59	1.34	-
	gpd	2.17	1.78	2.33	1.65	1.39	1.24
	MCMC	2.5	1.7	2.6	1.8	1.5	1.4

Table 2: table of threshold, extremal index and 50, 100 years return level, per month

Estimations of the extremal index is shown on table 2. The average cluster size are given by $1/\theta$. A unit extremal index indicates independence between the points and conversely a zero extremal index indicates dependence of all points. For PROD, the extremal indices are situated in the interval $[0.5, 0.7]$. Table 2 also displays the return level of 50 and 100 years, calculated with the point process, by fitting the GDP and using the Metropolis-Hastings algorithm. The formula for the return level x_m over m observations is

$$x_m = u + \frac{\sigma}{\xi}[(m\zeta_u)^\xi - 1] \tag{5}$$

where m is the product of the number of year and the number of points per month. For example, for January, $m = 50 \cdot 8 \cdot 31$ for the 50 years return level. u is the threshold and ζ is the probability that a data point is greater than the threshold. ζ is computed numerically. Some values on the table are

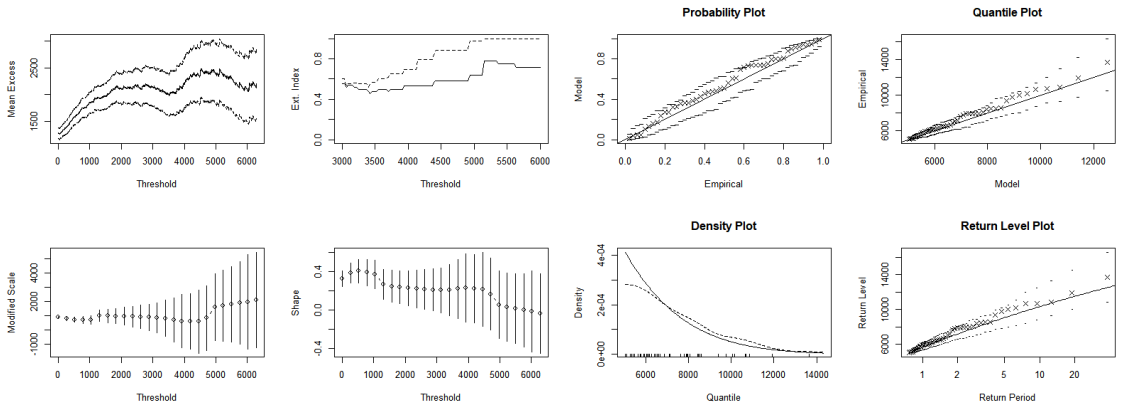

Figure 4: Mean residual life plot, extremal index plot, parameters stability plot on the left, Diagnostic plots on the right, for November

missing because of numerical instabilities. The values for MCMC are less precise (estimation from histograms) but are in general close to the return levels computed using the two other methods.

5 Annual Maxima analysis

Instead of considering monthly maxima over 37 years, that is, the monthly behaviour of the thunderstorms, an analysis of the annual maxima could be useful to expose the global trend of the extreme values over the years. A PROD maximum over one month is not necessary representative of a severe thunderstorm. Indeed major thunderstorms correspond to very high PROD value but over the year it depends on the season. For instance severe thunderstorms are rare in winter compare to spring. Annual maxima deal purely with the extremes thunderstorms. However, as shown on table 1, the parameters of the GEV distribution vary from month to month. Hence fitting the annual maxima does not give accurate prediction for a specific month, it only gives estimates for yearly extreme event without specifying in which month it is more likely to happen. The estimates and standard deviations of the fitted GEV using maximum likelihood can be found in table 3 and the diagnostic plots and profile log-likelihoods are shown on figure 5.

		max-lh	MCMC
η [m^3 s^{-3}]	$[10^3]$	$21.0_{0.71}$	$21.1_{0.83}$
τ [m^3 s^{-3}]	$[10^3]$	$3.74_{0.43}$	$4.23_{0.53}$
ξ	$[10^{-1}]$	$-0.58_{0.99}$	$-0.35_{0.12}$

Table 3: Table of parameters of the GEV for annual maxima, with the standard deviation in subscript. Estimated using maximum likelihood on the left, using Bayesian technique (MCMC) on the right.

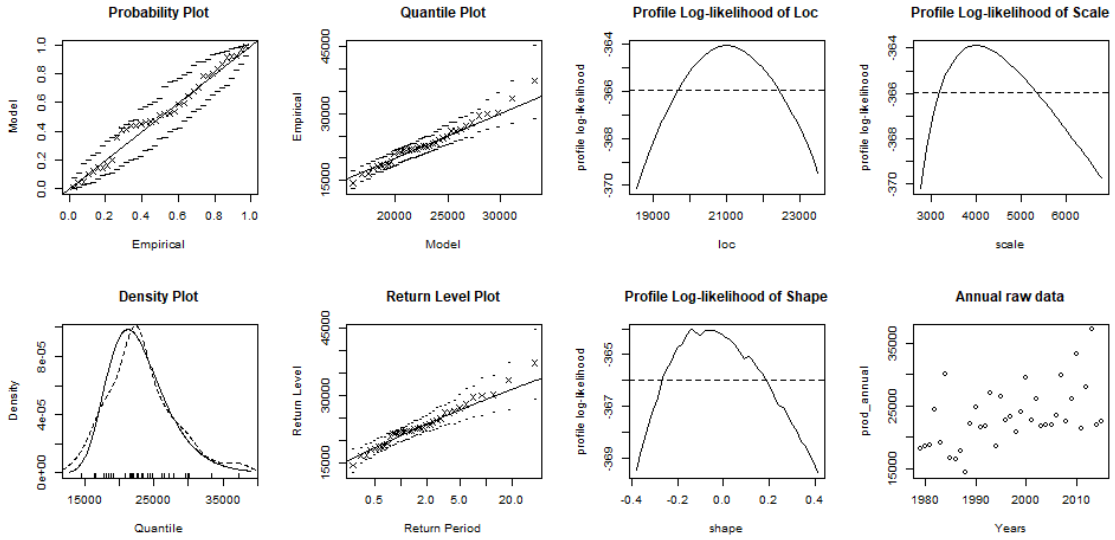

Figure 5: Diagnostic plots, profile log-likelihood and data for the annual maxima

6 Asymptotically dependence and bivariate models

The figure 6 shows the monthly maxima of CAPE and SRH and their asymptotic independence. Indeed, χ is estimated as smaller or close to zero and $\bar{\chi}$ is not close to the unity. Diagnostic plots of the logistic and the negative logistic models are shown on figure 7, all monthly maxima are considered together. The dependence (Pickands) function indicates that the two data sets are independent. To take the best model, the smallest Akaike information criterion (AIC) is considered:

$$AIC = 2\{dim(\theta) - l(\hat{\theta})\}, \tag{6}$$

where $l(\hat{\theta})$ is the log-likelihood evaluated on the estimated parameters. The other models do not fit and the AIC is 966.97 and 966.46 for the logistic and negative logistic models respectively, indicating that the negative logistic model is slightly better. To simplify the dependence analysis all monthly maxima were considered together. Indeed, it is very unlikely that the two data sets are dependent only for certain months. The bivariate analysis of each month separately was also done, indicating strong independence for each month.

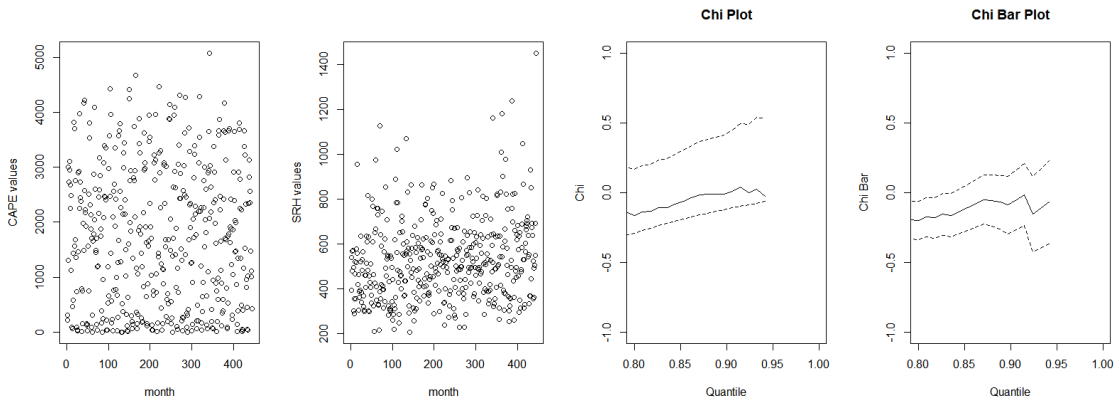

Figure 6: CAPE and SRH monthly maxima on the left, χ and $\bar{\chi}$ plots on the right, showing asymptotic independence.

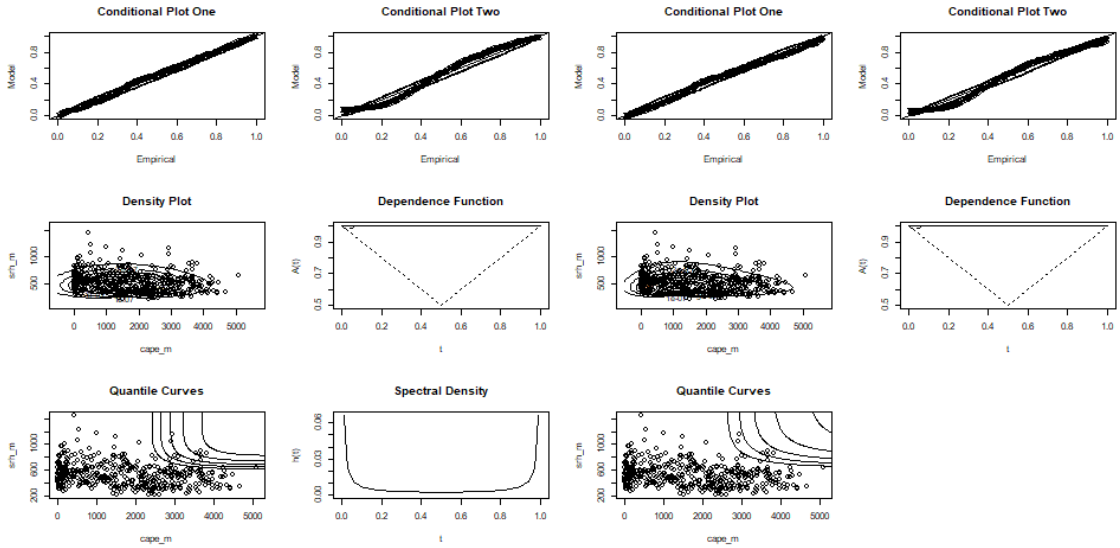

Figure 7: Diagnostic plots for the logistic and the negative logistic models

7 Analysis of the logarithm of the data points

According to the fits, taking the logarithm of PROD maxima change the type of the GEV. Indeed, for all months were it was possible to fit, namely from February to October, the shape parameter is negative. Hence every distribution of monthly maxima with each months considered separately is a reverse Weibull, meaning that a supremum exists probably for all months. Figure 8 displays the diagnostic plots and the profile log-likelihoods of the month of March where the logarithm of the points was taken, to make it comparable with figure 3. Taking the logarithm of a data set allows the study of the order of magnitude of the data points and set the values close to zero near minus infinity. Hence if a lot of values are small, they are more evenly space in a logarithm scale.

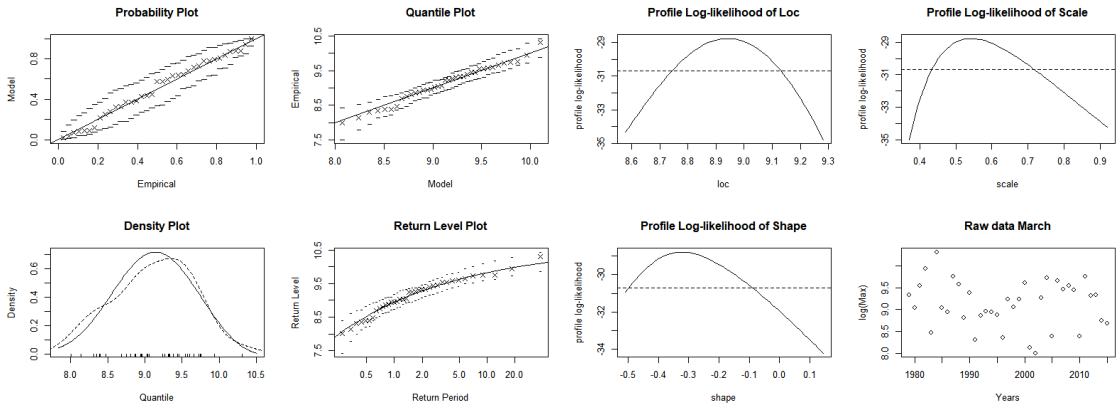

Figure 8: Diagnostic plots, profile log-likelihood and data for March, where the logarithm of the data points was taken.

8 Conclusion

The extremal analysis of PROD revealed that the fit of the GEV using maximum-likelihood, Bayesian methods or r-order statistics gives reasonable results, according to the diagnostic plots and the profile plots. The distribution is the reversed Weibull for the months of April to September and Frechet for the others, indicating a change in the extremal types twice a year. It was shown that it is unlikely that PROD is dependent of time or of ENSO, except maybe for June and August. Furthermore it is unlikely that CAPE and SRH are dependent and they are asymptotically independent, therefore it is likely that CAPE and SRH are also independent of ENSO. The computation of the 50 and 100 years return levels was made by two different approaches giving similar results. The annual maxima approaches is useful if one wishes to quantify annual risk, however it does not take into account that some values are more extreme at some specific months like in Spring. If time permits, it would be interesting to apply a better approach to find the optimal r for the r-order statistics like proposed in [3]. Moreover it would be interesting to further test the dependence of PROD and ENSO.

References

[1] [Online]. Available: https://en.wikipedia.org/wiki/Convective_available_potential_energy

[2] [Online]. Available: https://www.spc.noaa.gov/exper/mesoanalysis/help/help_srh1.html

[3] B. Bader, J. Yan, and X. Zhang, "Automated selection of r for the r largest order statistics approach with adjustment for sequential testing," *Statistics and Computing*, vol. 27, no. 6, pp. 1435–1451, 2017.

9 Annex

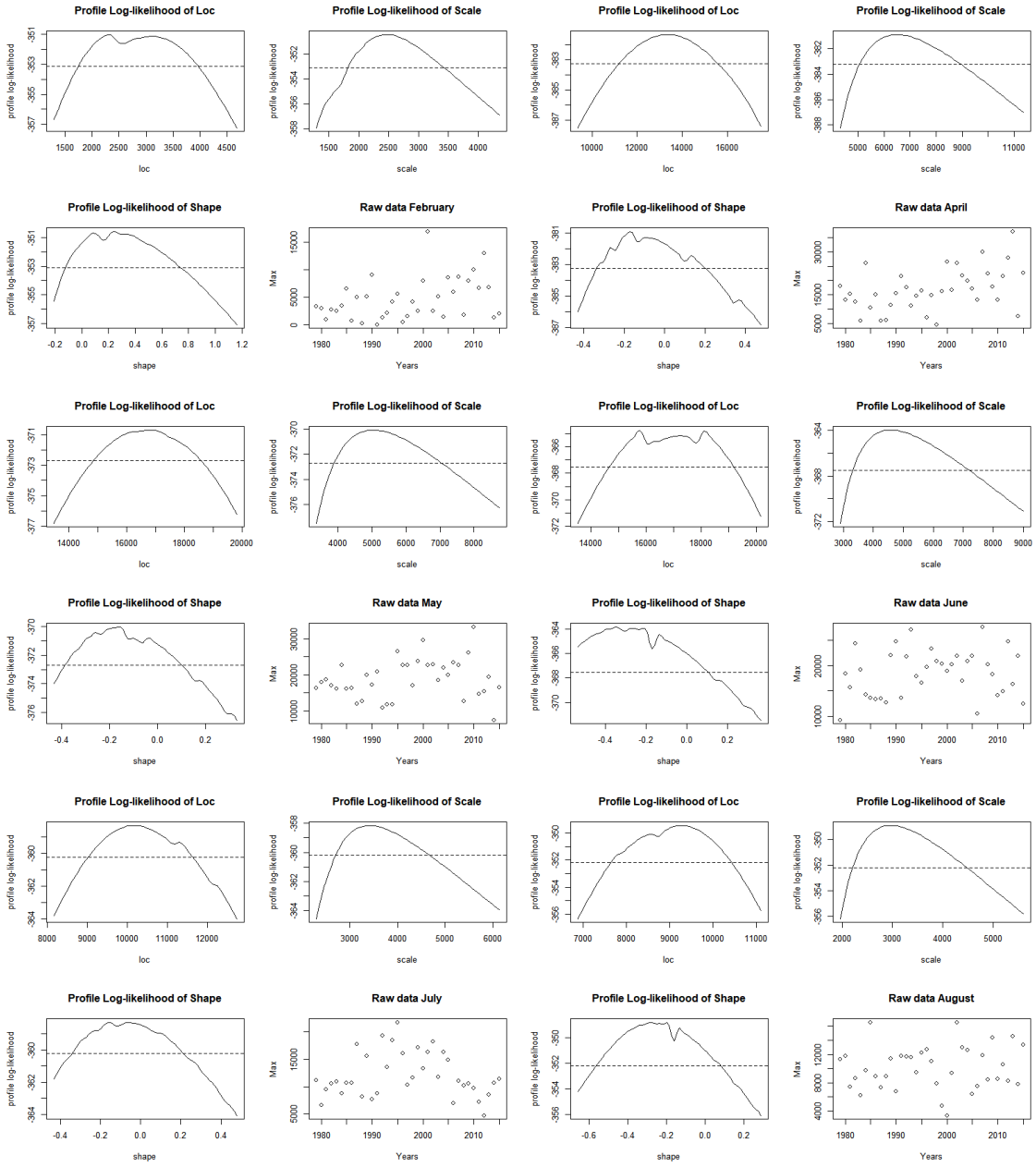

Figure 9: Diagnostic plots, profile log-likelihood and data for months from February to August, except March.

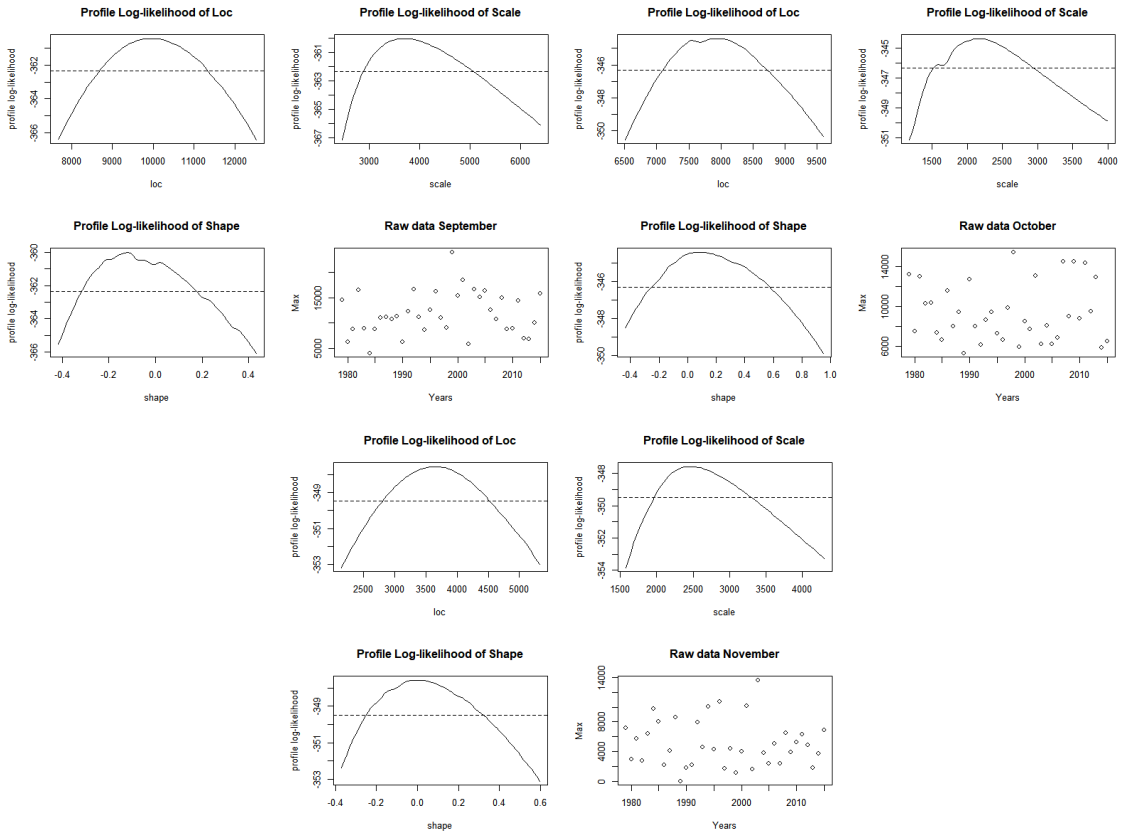

Figure 10: Diagnostic plots, profile log-likelihood and data for months from September to November.