PhD Progress

Marco Cucchi

October 24, 2018

Contents

1 Introduction

This document is a log-book for all the work done during my PhD project. All code used con be found on github repository `https://github.com/marco-cucchi/L96gev`.

2 EVT and the Lorenz-96 model

Aim of this work is to find EVT parameters for observables of the Lorenz-96 (L96) model, and compare them with the bounds provided in [1].

2.1 Lorenz-96 model simulations

As a first step, a number of independent simulations of the L96 model are performed. The L96 model is defined as follows. For $i = 1, ..., N$:

$$\frac{dx_i}{dt} = (x_{i+1} - x_{i-2})x_{i-1} - x_i + F \tag{1}$$

where it is assumed that $x_{-1} = x_{N-1}$, $x_0 = x_N$ and $x_{N+1} = x_1$. Here x_i is the state of the system on the i-th coordinate, and F is the forcing constant. For this set of simulations, values of F and N have been set to $F = 8$ and $N = 32$.
Integration has been conducted using 4-*th* order Runge-Kutta scheme, with integration step $dt = 10^{-2}$. Initial conditions for each simulation have been set equal to

$$x_i^0 = 8 + \epsilon, \quad \epsilon \sim U([-0.05, +0.05]) \tag{2}$$

Different levels of spatial aggregation, defined as $A = 32, 16, 8, 4, 2, 1$, have been considered:

- For $A = 32$ no aggregation is performed, and each value x_i is treated independently;
- For $A = 1$ all original N x_i values are spatially averaged into one single value $\overline{x}$ for each time-step;
- More in general, for $A = K$ the N spatial coordinates indicated by the index i are divided into K non-overlapping clusters c_j fixed in time, and corresponding x_i values belonging to the same cluster are averaged at each time-step.

As observable, the local energy of the system for different levels af aggregation

$$E_j = \frac{1}{2}x_j^2, \quad x_j = \begin{cases} x_i, & A = 32 \\ \overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i & A = 1 \\ \frac{1}{\#c_j}\sum_{i \in c_{j,K}} x_i & A = K \end{cases} \tag{3}$$

is considered.

In order to extract information on the statistics of extremes, very long simulations have to be performed. In order to find a good compromise between this requirement and the limited amount of disk space, a similar procedure to the one adopted in [2] has been followed: instead of keeping all values of each simulation, only block-maxima are retained, with block size $\Delta t = 0.5$. It is important to highlight that block-maxima are computed *after* aggregation (spatial average).

Following this procedure, for each simulation (initial condition) 6 different files are obtained, each corresponding to one particular aggregation level A: each of these files, then, contain A time-series of block-maxima, one for each of the A clusters.

Script: c003e11.

2.2 Statistics of Extreme Events

Parameters defining GEV distribution are estimated using three different approaches:

- Direct fit using *block-maxima* approach;
- Direct fit using *POT* approach (still not described hear);
- Method of *moments* described in [1].

Finally, estimates derived with these approaches are compared among them and with bounds related to attractor's dimensions described in [2].

2.2.1 Block-maxima approach

In this approach, each time-series for each different cluster of each simulation is fitted against GEV family of distribution separately. More specifically, for each cluster time-series belonging to a different simulation the following procedure is carried out:

1. Percentiles' orders p of interest are fixed (e.g. 0.99, 0.995, ...), and corresponding percentiles (thresholds) T_p are computed; [1]
2. Time-series is divided in n blocks, where $n = length(\text{time-series})(1-p)$;
3. Compute maxima for each block;
4. Fit GEVD family to the block-maxima series.

[1] This could be something to think upon; in this way I have (slightly?) different percentiles for different time-series in the same simulation and for different simulations. Is this right? The underlying assumption in this procedure should is that all time-series belonging to all simulations should come from the same distribution. So shouldn't the percentiles be the same for all of them?

The fit is performed with the R function `gevFit` from the package `fExtremes`, using MLE approach. As a result estimations of shape parameter ξ, location parameter μ and scale parameter σ are returned, with respective uncertainties as computed via MLE.
Location parameter μ is actually assigned the value T_p; the *absolute maximum* from each time-series is also kept; *modified scale parameter* σ^* is computed as

$$\sigma^* = \sigma - \xi T_p \tag{4}$$

Error on σ^* is estimated via propagation of error.
As explained in [3], in order to find a valid threshold value T_0 for excess to follow generalized Pareto distribution (and, consequently, GEV distribution), it is a good practice to plot ξ and σ^* against T_p and look for the value where both start to be approximately constant: that value is T_0.
Once parameters have been estimated for all clusters in a simulation, a single estimation of each parameter is saved as the average among all estimates.[2] Furthermore, the following parameters are estimated for each simulation:

- *scale parameter* σ is computed with inverse of equation 4, and relative error is computed via propagation of errors; [3]
- *upper end-point* is computed as[4]

  $$u\hat{e}p = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}} \tag{5}$$

  and the relative error is computed via propagation of errors.

Finally, for each different aggregation, ensemble averages of *shape* and *modified scale* among all simulations are computed.

2.2.2 Method of Moments

Following theory described in [1], we want to estimate *shape* and *scale* parameters using the following equations (Par 8.2.6 in [1]):

$$\xi_A^T = \frac{1}{2}\left(1 - \frac{(\langle \tilde{A}_1^T \rangle)^2}{\langle \tilde{A}_0^T \rangle \langle \tilde{A}_2^T \rangle - (\langle \tilde{A}_1^T \rangle)^2}\right) \tag{6}$$

$$\sigma_A^T = \frac{1}{2}\frac{\langle \tilde{A}_1^T \rangle \langle \tilde{A}_2^T \rangle}{\langle \tilde{A}_2^T \rangle \langle \tilde{A}_0^T \rangle - \langle \tilde{A}_1^T \rangle^2} \tag{7}$$

[2]The *absolute maximum* is also averaged, and this could be an error. The average of the *location parameter* μ is also a little disturbing, but this could be solved following reasoning in footnote **??**. Error computation should be checked.

[3]This sounds very stupid, since σ was originally estimated (but not saved) via MLE fit to GEVD.

[4]Find reference

where $A(x)$ is an observable of the system, T is a threshold value and

$$\langle \tilde{A}_n^T \rangle = \int \mu(dx) \Theta(A(x) - T)(A(x) - T)^n, \tag{8}$$

being Θ the Heaviside distribution. This results are exact in the limit for $T \to A_{max}$.
In order to perform this computation, the following procedure has been adopted. First, for each cluster time-series belonging to a different simulation:

1. Percentiles' orders p of interest are fixed, and corresponding percentile (thresholds) T_p are computed (footnote 1);

2. $\langle \tilde{A}_n^{T_p} \rangle$ for $n = 0, 1, 2$ are computed, using temporal average in place of ensemble average (assuming ergodicity).[5]

Once moments have been estimated for all clusters in a simulation, a single estimation of each moment is saved as the average among all estimates, and relative standard deviations are computed.
Using these estimates, *shape* parameter is computed through equation 6 and estimation of uncertainty is computed via propagation of error. Finally, for each different aggregation, ensemble averages among all simulations are computed.

2.2.3 Bounds to the *shape* parameter from the attractor's dimensions

We want to verify relation (8.2.15) in [1], which states that

$$(d_s + d_u + d_n)/2 \leq \delta \leq d_s + (d_u + d_n)/2, \tag{9}$$

where

- d_u is equal to the number of positive Lyapunov exponents of the system [4];

- d_n is equal to the number of zero Lyapunov exponents of the system, and in particular it is 1 for Axiom A systems [6];

- $d_s = n + \sum_{k=1}^{n} \lambda_k / |\lambda_{n+1}| - d_u - d_n$ [2], with λ_k denoting the Lyapunov exponents of the system, in a descending order, and n is such that $\sum_{k=1}^{n} \lambda_k$ is positive and $\sum_{k=1}^{n+1} \lambda_k$ is negative;

- $\xi = -1/\delta$;

- $\sigma = (A_{max} - T)/\delta$, with A_{max} and T denoting the maximum observed value of the observable[7] and the threshold value.

[5] No standard deviation has been computed at this stage!
[6] We are taking this for true in our system
[7] or the *upper end point*?

Lyapunov exponents have been computed using Benettin algorithm with QR decomposition. Bounds have been computed and averaged over 50 iterations (simulations).
Script:

- Lyapunov exponents computation: 136afd4
- Average bounds computation:

2.3 Statistics of Extreme Events: Corrections and Results

In this section results of the analyses reported in Sec. 2.2 are described, after issues highlighted in the footnotes 1,2,3.
Script:

- quantiles computation: 6880d20

2.3.1 Block-maxima approach

The following corrections have been applied:

- Percentiles are computed once, concatenating the first 80 simulations of the first clusters for each aggregation;
- Shape, scale and location parameters from fit procedures are saved for each cluster in each simulation. Averages and computation of derived parameters come after;

Results are shown in Fig. 1 and 2.
Script:

- fit: 240d6d0
- parameters derivation and plots: 97925ef

2.3.2 Method of Moments

Results are shown in Fig. 3 and 4.
Script:

- moments computation: 240d6d0
- parameters derivation and plots: 97925ef

3 LRT and the Lorenz-96 model

Aim of this work is to study the possibility of applying LRT to extremes of observables of the Lorenz-96 model

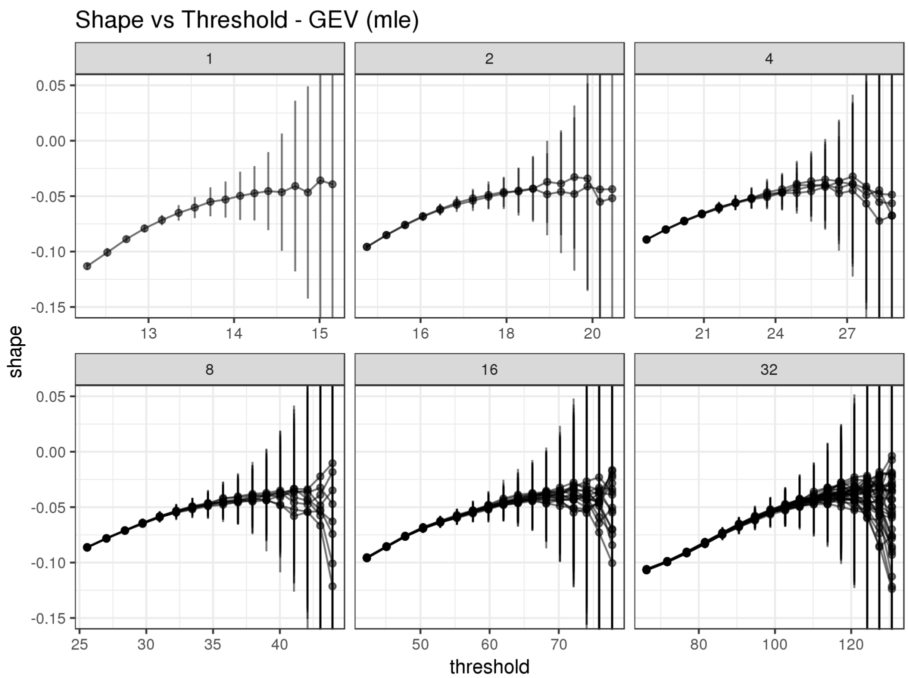

Figure 1: Ensemble average of *shape* parameter over 92 simulations. Each cluster is treated separately.

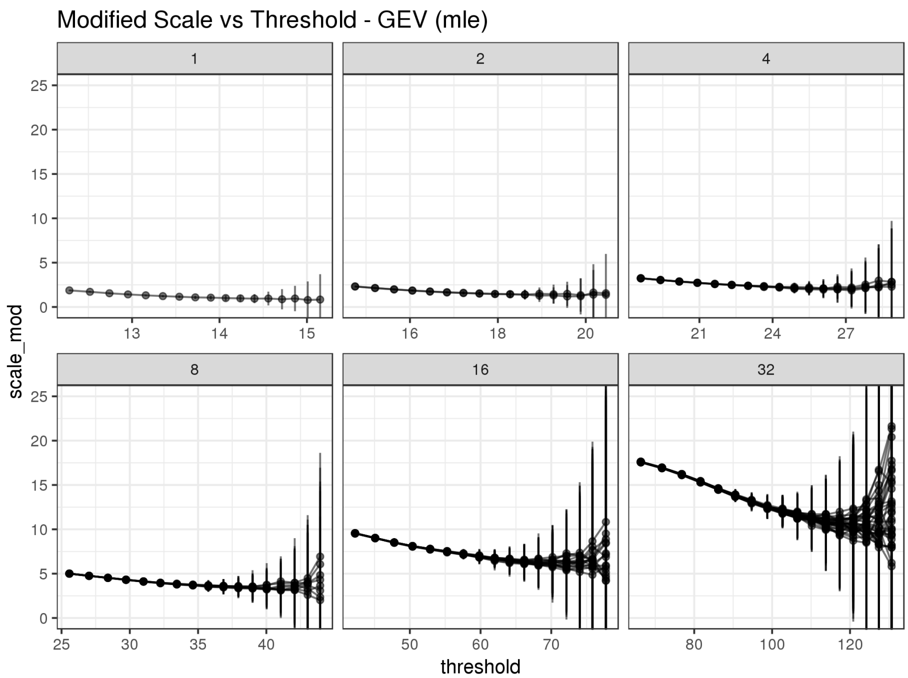

Figure 2: Ensemble average of *modified scale* parameter over 92 simulations. Each cluster is treated separately.

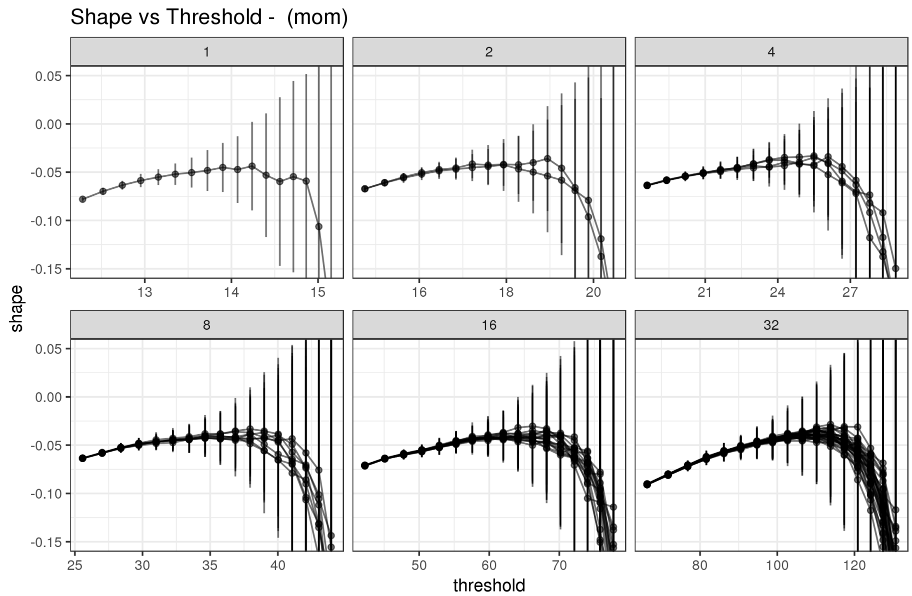

Figure 3: Ensemble average of *shape* parameter over 92 simulations. Each cluster is treated separately.

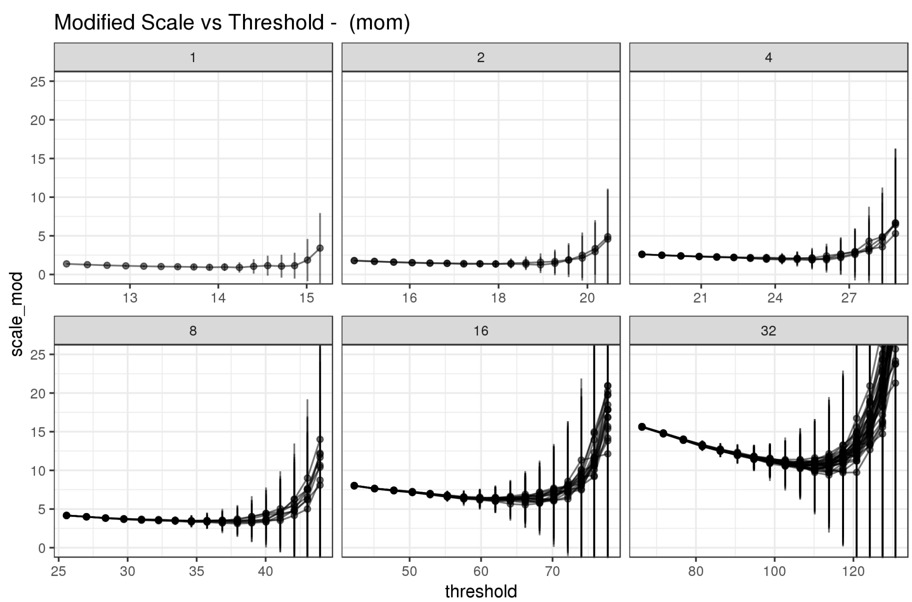

Figure 4: Ensemble average of *modified scale* parameter over 92 simulations. Each cluster is treated separately.

References

[1] V. Lucarini, *Extremes and Recurrence in Dynamical Systems.* Wiley, 2016.

[2] M. Galfi, T. Bòdai, and V. Lucarini, "Convergence of extreme value statistics in a two-layer quasi-geostrophic atmospheric model," *Complexity*, 2017.

[3] S. Coles, *An Introduction to Statistical Modeling of Extreme Values.* Springer, 2001.

[4] E. Ott, *Chaos in Dynamical Systems.* Cambridge University Press, 2002.