Learning Extremes with evtGAN

Younes Boulaguiem, PhD
Postdoctoral Researcher in Statistics

University of Geneva – Faculty of Medicine
Clinical Research Center, HUG

A generative framework for learning and simulating spatial extremes.
Delivered June 2020

View Publication

Context and motivation

- Understanding **environmental extremes** events such as **floods**, **heatwaves** and **heavy rainfalls** is of paramount importance since they often lead to **severe socio-economical impacts**, more particularly compound events, which are rare events characterized by the concurrent occurrence of multiple and possibly interdependent hazards.
- Every few years, the **Intergovernmental Panel** on Climate Change [Meehd et al.] compares the state-of-the-art models and derives recommendations in their scientific report. In the more recent versions, a whole chapter is devoted to climate extremes.
- The characterisation of the **underlying probabilistic structure** of rare events is a **difficult task**, and the development of models that are able to **extrapolate beyond the bulk** of a distribution is key.
- In this talk I will briefly present the approaches that exist to characterise the extremes, their advantages and limitations, and how we can take advantage of their strengths in the conception of a new statistical methodology that proves to be useful and competitive.

A glimpse on the data

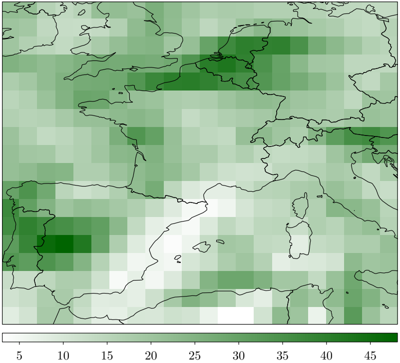


Figure: **Component-wise maxima** of daily precipitations (mm/day) over a period of one year. Each **pixel represents a location** in the map where the measurements have been taken consistently.

- Our dataset contains the daily precipitations for **396 locations** over Europe mainly for 2'000 years (they are the results of simulations).

A glimpse on the data

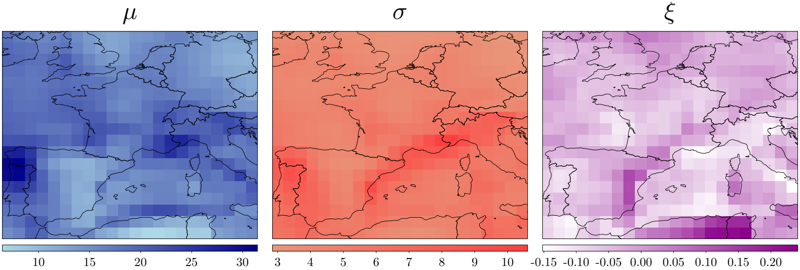


Figure: GEV parameter estimates. On the left the mean parameter μ, in the middle the scale parameter σ and on the right the shape parameter ξ.

- The **shape parameter** is the most important parameter as it indicates whether a margin j is **heavy-tailed** ($\xi_j > 0$), **light-tailed** ($\xi_j = 0$) or has a **finite upper end-point** ($\xi_j < 0$).

Classical approach

- **Observational datasets** represent **only one of the many** possible realisations of the climate system and thus are often **ill-suited** in characterising the probability of **rare events**.
- As a consequence, the main route for the analysis of our future climate relies on **large scale simulations** of climate models, which are based on the thorough analysis of the physical processes of the atmosphere.
- However, this entails a **huge computational cost** (several years) when the main interest is in extremes. Indeed, we need to observe sufficiently many rare events to perform meaningful inference.

Statistical approach

- **Extreme value theory** is accurate in estimating **the marginal distribution** and provides mathematically justified models for the tail region of a multivariate distribution, enabling **extrapolation beyond the range of the data** and accurate estimation of the small probabilities of rare events.
- However, when it comes to estimating the dependence structure, the models remain either **simplistic** or **overparametrised** as they require very **strong assumptions**, such as **isotropic** dependence and **stationarity**, when most real life large spatial domains exhibit **non-stationarities**.
- A state-of-the-art model used in this context is the **Brown-Resnick** [Davison et al., 2012] which we will use as the standard for comparison in the upcoming results.
- In large dimensions, they can also be **computationally intensive** because of complicated likelihoods.

Machine learning approach

- Models like the **Generative Adversarial networks** (GANs) seem to be quite competitive to learn probability density functions.
- Quite **flexible** and well suited for **complex** and **non-stationary** data.
- **Computationally cheap**! (relatively speaking)
- **BUT**, loss functions are designed to **predict in the bulk** of the distribution, and it is difficult to construct approaches with a good performance **outside the range** of the training data. That is why these models have generally been discarded in the extreme value literature.

In sum

- Limitations:
  - Classical approach: computationally costly (several years).
  - Stat. approach: strong assumptions required to accurately model the dependence structure, computationally costly in high dimensions.
  - ML approach: weak performance outside the space of training data.
- Strengths:
  - Classical approach: accurate.
  - Stat. approach: accurately estimates the marginal structure using Generalised Extreme Value (GEV) distribution fit (Fisher-Tippet-Gnedenko theorem).
  - ML approach: accurately estimates the dependence structure.

evtGAN

- Our main idea is to **disentangle** the learning of the **dependence** structure and the **marginal** distributions using a **copula-based** approach and providing for each one the **most appropriate solution**.
- As such, we combine the asymptotic theory of extremes with the flexibility of GANs, to build a more **efficient statistical model** that would serve as an **emulator** specifically designed for extreme events, able to reproduce the **spatial tail dependencies and to extrapolate outside the training space**, starting from as few as 30 annual maxima!
- This is what we have called **evtGAN**.

Procedure

- Let $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{id})$ be the component-wise maxima computed from the original data and $i = 1, \cdots, n$.

Algorithm

1. On the basis of n training data points $\{\mathbf{X_1} = (X_{1,1}, \ldots, X_{1,d}), \ldots, \mathbf{X_n} = (X_{n,1}, \ldots, X_{n,d})\}$ compute univariate GEV parameter estimates for each location of the map, $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_d)$, $\hat{\sigma} = (\hat{\sigma}_1, \ldots, \hat{\sigma}_d)$ and $\hat{\xi} = (\hat{\xi}_1, \ldots, \hat{\xi}_d)$.
2. Normalize the margins of each location in the training dataset through the transformation $\mathbf{X_i}^* = (X^*_{i,1}, \ldots, X^*_{i,d}) = (\hat{F}_1(X_{i,1}), \ldots, \hat{F}_d(X_{i,d}))$, $i = 1, \ldots, n$, where $\hat{F}_j(\cdot)$, $j = 1, \ldots, d$, is the empirical cumulative distribution computed for the $j-$th location.
3. Train a GAN on $\{\mathbf{X_1}^*, \ldots, \mathbf{X_n}^*\}$
4. Generate m data points from G and normalize empirically its margins to uniform distributed margins. Denote these data points with $\mathbf{G}_1, \ldots, \mathbf{G}_m$.
5. Normalize back to the original scale the data points $\mathbf{G}_1, \ldots, \mathbf{G}_m$ using the reverse GEV transformation with the parameters estimated in step 1.

Metrics

- In order to measure the quality of the multivariate extrapolation, we consider the following metrics:

1. **The extremal coefficients**:
    - Given two locations in the map with Fréchet marginal distributions, the extremal coefficient $\theta \in (1, 2)$ **measures the strength of the dependence** between them as it appears in the following expression:

$$\lim_{x \to +\infty} \Pr \left(X_2 > x | X_1 > x \right) = 2 - \theta.$$

    - The extremal coefficient can be interpreted in the light of the limiting **probability of one variable being large given the other is large**.
    - In particular, X_1 and X_2 are said to be asymptotically **independent if** $\theta = 2$, while they are **perfectly dependent for** $\theta = 1$.

Metrics

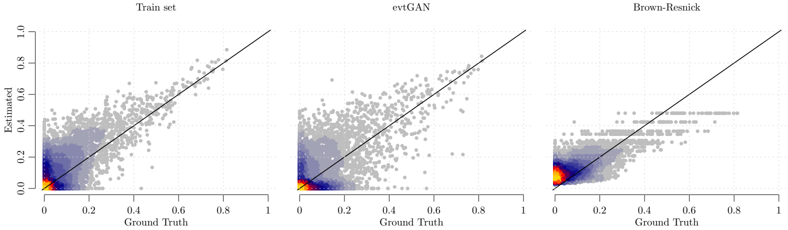
2. **The polar representation**:
   - In the bivariate case, the variable (X_1, X_2) can be expressed in the polar representation:

   $$(R, W) = \left(X_1 + X_2, \frac{\exp(X_1)}{\exp(X_1) + \exp(X_2)} \right) \in [0, \mathbb{R})x(0, 1).$$

     R is what is called the radial component, and W is called the angular component.
   - Under certain conditions, it can be shown that **the angular component has a well defined distribution**. But in general, **conditionally on** R **being large**, observations for which $W \approx 0$ or $W \approx 1$ correspond to data points for which either X_1 or X_2 is extreme, in which case the variables are **independent**. Whereas when $W \approx 1/2$, both X_1 and X_2 are extremes, in which case the variables are **perfectly dependent**.
3. **Empirical bivariate distribution**: this will allow us to visualise the **extrapolation** performances of our models outside the training space.

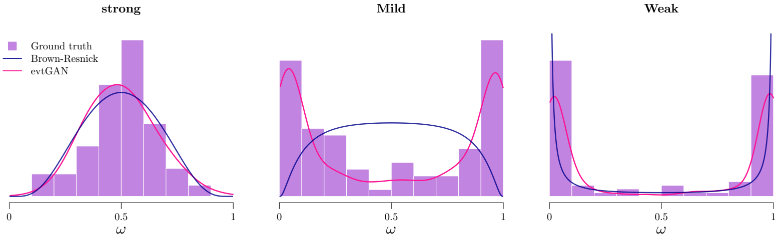
- We perform a train-test split on the data, with $n_{train} = 50$ and $n_{test} = 1950$. The **test set** will represent our **ground truth**.
- We will compare the results of **evtGAN** with the ones obtained with the **Brown-Resnick** model and the **ground truth**.

Results



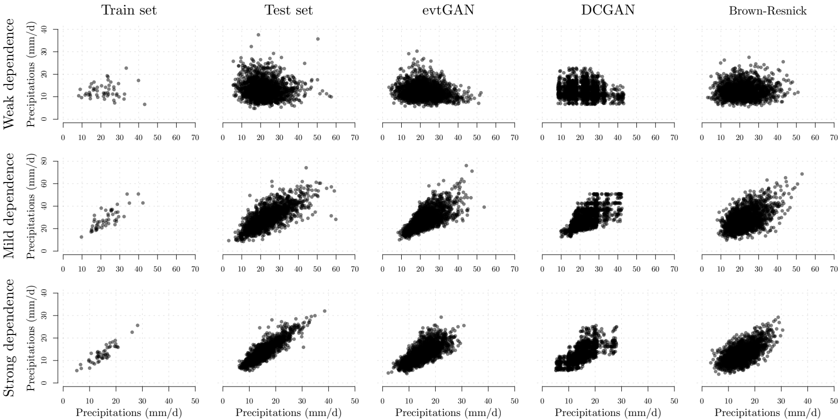
- What is displayed above is $2 - \theta$. The **extremal coefficients** computed for the **same pairs of locations** on each of the train set, 10'000 observations generated from evtGAN (trained on the train set), and analytically from the Brown-Resnick model (fitted on the train set), are **plotted against** the extremal coefficients computed on the **test set** (our ground truth).
- The **Brown-Resnick** model uses a **parametric semivariogram** based on the **euclidean distance** to describe the spatial continuity of the data and their **interdependence** (isotropic dependence assumption). Meaning that all the pairs of locations with the **same distance** will yield the **same value** for the extremal coefficients, which explains the **stair-like behaviour**.
- Another limitation of the **Brown-Resnick** is that it **always assumes dependence**, which is why the extremal coefficients are always **smaller than 2**. This is a huge constraint in this context.

Results



- Notice how in the case of mild dependence the **Brown-Resnick** completely **misspecifies the underlying spectral distribution**.

Results



Train set
Test set
evtGAN
DCGAN
Brown-Resnick
Weak dependence
Mild dependence
Strong dependence
Precipitations (mm/d)

Conclusions

- Understanding and evaluating the risk associated with extreme events is of major importance to our society.
- In the case of climate science, using **large ensemble simulations** from physical models could **infeasible** given their high computational cost and the amount of data required to include sufficient rare events. Similarly **historical records** are usually **too short** for a meaningful analysis of extremes.
- **evtGAN combines the best of machine learning and extreme value theory** and offers, as far as we know, the **best computational efficiency**.
- With its **easy implementation**, it offers the scientific community a ready-to-use **emulator specifically designed for extremes**, and is able to reproduce the **univariate maxima distributions** and **tail dependencies** with **great accuracy** and very **few training examples**, with applications ranging from climate studies to epidemic diffusion analyses and finance.
- Besides, this work represents a glance on the potentiality of this methodology. The results could easily and continuously be improved with the evolution of more sophisticated deep learning architectures and more efficient statistical estimations, and further extended to include the temporal dimension.

References

Jan Beirlant, Yurtal Goegebeur, Johan Segers, and Jozef Teugels. *Statistics of Extremes: Theory and Applications*. Wiley, 2004.

Li Cheng, Reimund P Rötter, et al. Hdr: A hybrid deep-learning and regularized statistical model for estimating extremes under climate change. *Environmental Modelling & Software*, 165:105732, 2023.

Stuart Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer, 2001.

Anthony C Davison, Simone A Padoan, and Mathieu Ribatet. Statistical modeling of spatial extremes. *Statistical Science*, 27(2):161–186, 2012.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Jana Sillmann, Viatcheslav V Kharin, Francis W Zwiers, Xuebin Zhang, and Don Bronaugh. Extreme weather and climate events in a changing climate. *Nature Reviews Earth & Environment*, 1(1):1–15, 2020.

Jennifer L Wadsworth and Jonathan A Tawn. Deep learning for the estimation of tail probabilities. *Extremes*, 25:1–28, 2022.