University of Glasgow

Statistical Modelling of Environmental Extremes

– Session 1 –

Dr Daniela Castro-Camilo

SCAN FOR COURSE MATERIAL

Institut Henri Poincaré, Sorbonne University – March 2022

Plan for today

Plan for today

- [9:00-10:00] **Session 1:** short introduction to the modelling of environmental extremes. Recap of classical modelling techniques and gentle introduction to INLA.
  - No R/RStudio required.
  - Documents needed: SMEE_IHP2022_Session1HO.pdf.
- [10:05-10:30] **Session 2 and lab:** we learn how to tackle trends and seasonalities in block-maxima data.
  - R/RStudio required.
  - Documents needed: SMEE_IHP2022_Session2HO.pdf and Practical2Lab.pdf.
- [11:00-12:30] Session 2 and lab cont'd.

What is this session about?

What is this session about?

1. Motivates the need to use statistics of extremes to study environmental processes.

2. Provides a (non-exhaustive) summary of common modelling challenges when analysing environmental extremes, possible courses of action in terms of modelling and inference techniques.

3. Reviews (very quickly) some useful classical (and probably well-known) statistical techniques.

4. Introduces the integrated nested Laplace approximation (INLA) to model threshold exceedances using latent Gaussian models.

Note: In 4, I will show an example code to fit a model using R-INLA. This is for illustration purposes only (meaning, you are not expected to run the code in your computer). We will run INLA fits in Practical Session 4.

Introduction to the Statistical modelling of environmental extremes

Why Statistics of Extremes is important in environmental sciences?

- In much of the environmental sciences and particularly in risk assessment, extremes events play a crucial role due to the severity of the consequences for human life and the ecosystem. Think of earthquakes, tsunamis, fires, floods, hurricanes, etc.

- Extreme value analysis is key in the implementation of environmental policies, since they usually involve the development of standard for environmental variables by government agencies.

- These standards usually involve the selection of a high (or low) threshold, which in turn define an extreme event (e.g., whether the annual maxima of PM10 exceeds 50 $\mu g/m^3$).

Challenges in (extremes of) environmental variables

- Environmental variables usually exhibit different types of non-stationarity, such as trends, seasonality or in general non-identical distributions across observations.

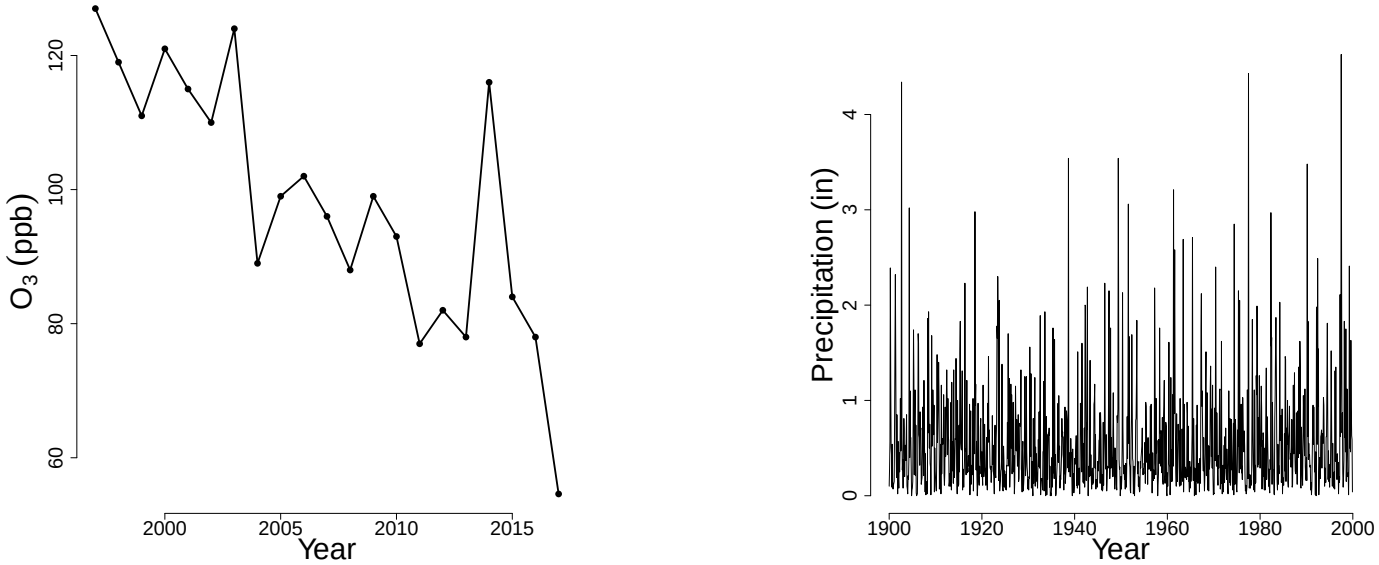


Figure 1: Left: Annual Ozone concentrations in Santiago, Chile. Right: Daily precipitations in Forth Collins, Colorado, US.

- For the case of variables observed over space, they might also exhibit different levels of dependence across locations (spatial dependence)
- Typical studies on environmental variables focus on the mean (and sometimes on the standard deviation), not the extremes. Therefore, the removal of trends, seasonality and the process of making data stationary are done at the mean level, and model residuals are obtained.

Challenges in (extremes of) environmental variables

- A traditional approach would be to apply an extreme value (EV) analysis to the residuals. But this approach has many red flags.
  - Removing trend/seasonality in the overall mean would not necessarily eliminate trends/seasonalities in the extreme values.
  - This is a two-step approach where uncertainty needs to be propagated (i.e., uncertainty from the trend/seasonal model needs to be passed on to the uncertainty of the EV analysis). In practice, uncertainty propagation is usually ignored .
- Bottom line is, in general, there is no reason to believe that the form of trends/cycles/non-stationarity on extremes is identical to that for the mean.

  Example: while analysing ground-level ozone, Smith (1989) found no trend in the overall levels of the series, but a marked downward trend in the extremes.
- In this course, we will take an EV approach, where the parameters of the EV distribution are "expanded" to account for challenging behaviours of the extremal data.

Here is a (non-exhaustive) summary of common modelling challenges when analysing environmental extremes, possible courses of action and inference techniques.

I. Modelling

1. Modelling extremes of a single environmental variable

- Such as: temperature, precipitation, pollutant concentrations, etc.
- What to do with trends/seasonality/non-stationarity: "expand" the parameters of the EV distribution, i.e., use linear and flexible regression to accommodate for trends, seasonality, influence of covariates, etc. [we are doing this in Practicals 2 and 3]
- What to do with spatial patterns
  - Environmental variables are usually measured over space using on-site stations, satellite data, etc.
  - Close locations tend to share similarities: "everything is usually related to all else but those which are near to each other are more related when compared to those that are further away" (first law of geography. Waldo Tobler, 1970).
  - That rule also applies to extremes of environmental variables.
  - Spatial extreme focuses on joint probabilities of extreme events across a continuous index space.

Statistical tools for extremes of environmental variables

- What to do with spatial patterns (cont'ed)
  - There are different ways for modelling spatial extremes:

    - Using a continuous version of the extremal types theorem that gives rises to max-stable processes [we are not doing this; see, e.g., Davison et al., (2012)]

    - Using more standard methodologies for spatial statistics, based on latent spatial process models. Here, dependence across locations is induced by a latent process [we are doing this for a special case in Practical 3.]

    - Modelling extremes location-wise and then use smoothing techniques (such as krigging) to integrate the results and extrapolate to locations with no data (in general, not a good idea because it is a two-step approach that smooths out the extremes) [we are not doing this; see, e.g., Smith (2013)].

2. Modelling joint extremes of two or more environmental variables with no spatial component
   - Such as: wave and surge hight, wind speed and wave height, wind speed and wind gust, etc.
   - Focuses on joint probabilities of extreme events, which implies modelling the dependence between these environmental extremes.
   - Different ways to model dependence: parametrically [we are doing this in Practical 5], semi-parametrically [we are not doing this; see, e.g., Castro-Camilo et al., (2018)].
3. Modelling joint extremes of two or more environmental variables with spatial component
   - Focuses on the way the dependence of these environmental extremes changes with space. [we are not doing this; see, e.g., Genton et al. (2015), Shooter et al., (2022) and references therein]

II. Inference

Inference for EV applications should be tailored to the model at hand. In recent years, Bayesian inference has become widely popular in EV analysis partially due to need to use flexible models that account for different sources of non-stationarity in the data or different levels of complexity.

- Maximum likelihood-based inference: classical inference technique that uses numerical optimisation techniques. Complications might arise due to the form of the likelihood. Approximate methods to tackle these issues are available. [we are using this]
- Bayesian inference
  - Allows for parameters to be random variables and therefore described by a distribution. This distribution get updated as data becomes available.
  - The focus is in the posterior distribution, which is a combination of the information from data (likelihood) and the prior knowledge of the parameters (prior distribution).
  - The posterior distribution (or features of the distribution such as the mode) usually involve intractable integrals, and numerical techniques have been develop to approximate them. Some examples are:
    - Monte Carlo methods (MC integration, importance sampling, rejection sampling)
    - Markov chain Monte Carlo (MCMC) methods
    - Laplace approximation
    - Numerical integration
    - Laplace approximation + numerical integration = integrated nested Laplace approximation INLA [we are using this in Practical 4]
  - Bayesian inference is natural (although not necessarily easy!) for hierarchical models.

Recap of useful modelling techniques

Recap of useful modelling techniques

- Standard linear model (LM): a linear model with fixed effects ($\boldsymbol{\beta}$) and normal response.

$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

$$\mu_i = \sum_{j=1}^{J} \beta_j x_{ij}$$

  $(x_{i1}, \ldots, x_{ij})$ are covariates.

- Linear mixed model (LMM): a LM with fixed ($\boldsymbol{\beta}$) and random ($\mathbf{b}$) effects. Allow for some degree of non-linearity in the model structure.

$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

$$\mu_i = \sum_{j=1}^{J} \beta_j x_{ij} + \sum_{k=1}^{K} b_l x_{ik}$$

$$\mathbf{b} = (b_1, \ldots, b_L) \sim \mathcal{N}(0, \boldsymbol{\psi})$$

- Generalised linear model (GLM): allow for response distributions other than normal.

$$Y_i \sim \mathrm{EF}(\mu_i, \phi)$$

$$\mu_i = \mathrm{E}(Y_i)$$

$$g(\mu_i) = \sum_{j=1}^{J} \beta_j x_{ij} := \eta_i$$

  EF is any distribution in the exponential family, ϕ is the scale parameter, g is the link function ($g(z) = z$ for LM and LMM) and η_i is the linear predictor.

Recap of useful modelling techniques

- Generalised linear mixed model (GLMM): allow for response distributions other than normal and some degree of non-linearity in the model structure.

$$Y_i \sim \text{EF}(\mu_i, \phi)$$

$$\mu_i = \text{E}(Y_i)$$

$$g(\mu_i) = \eta_i = \sum_{j=1}^{J} \beta_j x_{ij} + \sum_{k=1}^{K} b_k x_{ik}$$

$$\mathbf{b} = (b_1, \ldots, b_K) \sim \mathcal{N}(0, \psi)$$

- Additive models: a LM where the linear predictor (i.e., the mean) is expressed as a sum of smooth functions of covariates.

$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

$$\mu_i = \alpha + \sum_{l=1}^{L} f_l(x_{li})$$

- Generalised additive models (GAMs): a GLM where the linear predictor is expressed as a sum of smooth functions of covariates.

$$Y_i \sim \text{EF}(\mu_i, \phi)$$

$$\mu_i = \text{E}(Y_i)$$

$$g(\mu_i) = \eta_i = \alpha + \sum_{l=1}^{L} f_l(x_{li})$$

- Generalised additive mixed models (GAMMs): a GAM incorporating fixed and random effects (or a GLMM where the linear predictor also depends on some smooth functions of covariates).

$$Y_i \sim \text{EF}(\mu_i, \phi)$$
$$\mu_i = \text{E}(Y_i)$$
$$g(\mu_i) = \eta_i = \sum_{j=1}^{J} \beta_j x_{ij} + \sum_{k=1}^{K} b_l x_{ik} + \sum_{l=1}^{L} f_l(x_{li})$$

Notes

- For GAMs, the smooth functions are ***represented*** using basis (known) functions. For example, suppose that f_1 is believed to be a 4^{th} order polynomial. A basis for this space is $b_1(x) = 1, b_2(x) = x, b_3(x) = x^2, b_4(x) = v^3, b_5(x) = x^4$. Then, f_1 can be *represented* as

$$f_1(x) = \sum_{j=1}^{5} b_j(x)\tilde{\beta}_j = \tilde{\beta}_1 + x\tilde{\beta}_2 + x^2\tilde{\beta}_3 + x^3\tilde{\beta}_4 + x^4\tilde{\beta}_5,$$

  where $\tilde{\beta}_j$ are unknown coefficients (and therefore need to be estimated).

How this recap connects to the following sessions

- In Practical sessions 2, 3 and 4 we will focus on maximum likelihood inference for generalised linear models (GLM) and generalised additive (mixed) models (GAMMs) for the generalised extreme value (GEV) distribution and the generalised Pareto distribution (GPD).
- These model will help us tackling common challenges in environmental extremes, such as trends, seasonalities and the influence of covariates.
- GLM/GAMMs are also very useful if we want to include spatial patterns for extremes observed over space (e.g., temperature, precipitation, pollution at different stations).
- There are computational tools to fit GLM/GAMs using maximum likelihood (e.g., the very useful `extRemes` and `evgam` R packages, to mention a few.)
- But here we will focus on flexible modelling of spatial extremes using Bayesian inference.
- In the next section we will learn one fast and accurate way to fit extreme-value models with GAMM structure in a Bayesian framework.

Introduction to INLA

Introduction to INLA - Latent Gaussian models for threshold exceedances

- The integrated nested Laplace approximation is an inference technique that allows us to fit a class of complex Bayesian models (called *latent Gaussian models*) much faster (and accurately) compared to other Bayesian approximation methods, such as MCMC.
- A latent Gaussian model can be described using a GAMM structure as before. For example, if data $\mathbf{y}$ is assumed to follow a GPD, then

$$Y_i \sim \text{GPD}(q_\alpha(i), \xi), \quad \text{where } q_\alpha \text{ is the } \alpha \text{ quantile}$$

$$\log(q_\alpha(i)) = \eta(i) = \alpha + \sum_{m=1}^{M} \beta_m x_m(i) + \sum_{k=1}^{K} f_k(z_k(i)),$$

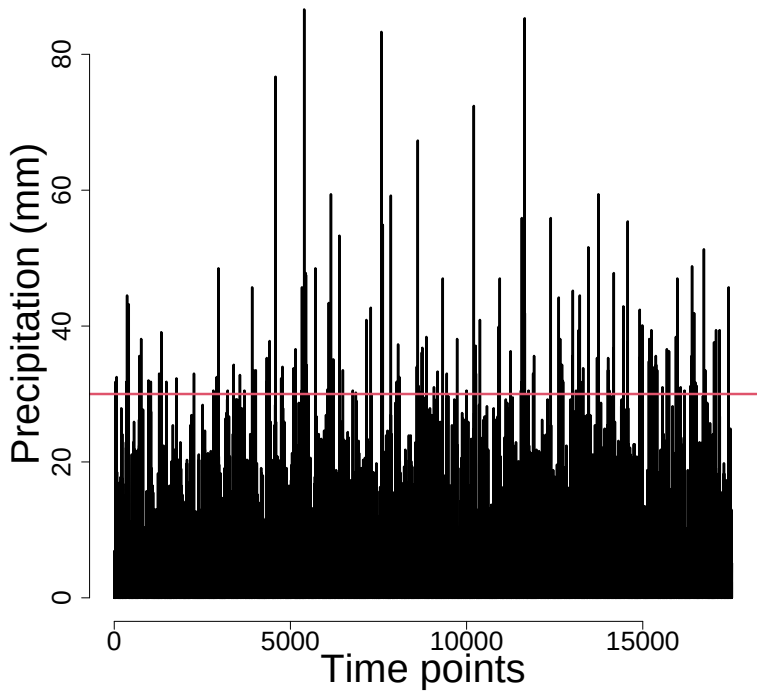
- We can see that the linear predictor is linked to the α-quantile (recall that in GLM/GLMM/GAMMs, it is usually linked to the mean ⏩).
- For instance, if $\alpha = 0.5$, then the linear predictor describes the effect of covariates over the median, and the GP is parametrised in terms of the median and the shape parameter.
- The usual GPD scale parameter can be obtained by

$$\sigma = \frac{\xi q_\alpha}{(1-\alpha)^{-\xi} - 1}$$

- We could also have a similar model for block-maxima, but in this course, we will focus on the threshold exceedances model only.

Introduction to INLA – Example in R

- Consider the data $\texttt{rain.txt}$ which contains daily rainfall accumulations (in mm) at a location in England over the period 1914-1962.



- We are interested in modelling excesses over 30 mm.

Introduction to INLA – Example in R

- We assume the following model

$$y(t) \sim \text{GP}(q_{0.5}(t), \xi)$$
$$\eta(t) = \log\{q_{0.5}(t)\} = \alpha + \beta t, \quad t = 1, 2, \ldots$$

- The following code fits the model in INLA and provides a summary of the fit.

```
library(INLA)
inla.data = data.frame(time = 1:length(rain),
                       y = rain,
                       intercept = 1)
inla.data = inla.data[inla.data$y>30, ]
formula   = y ~ -1 + intercept + time
crl.fam   = list(control.link = list(quantile = 0.5))
inlafit   = inla(formula,
                 family = 'gp',
                 data = inla.data,
                 control.family = crl.fam)
summary(inlafit)
```

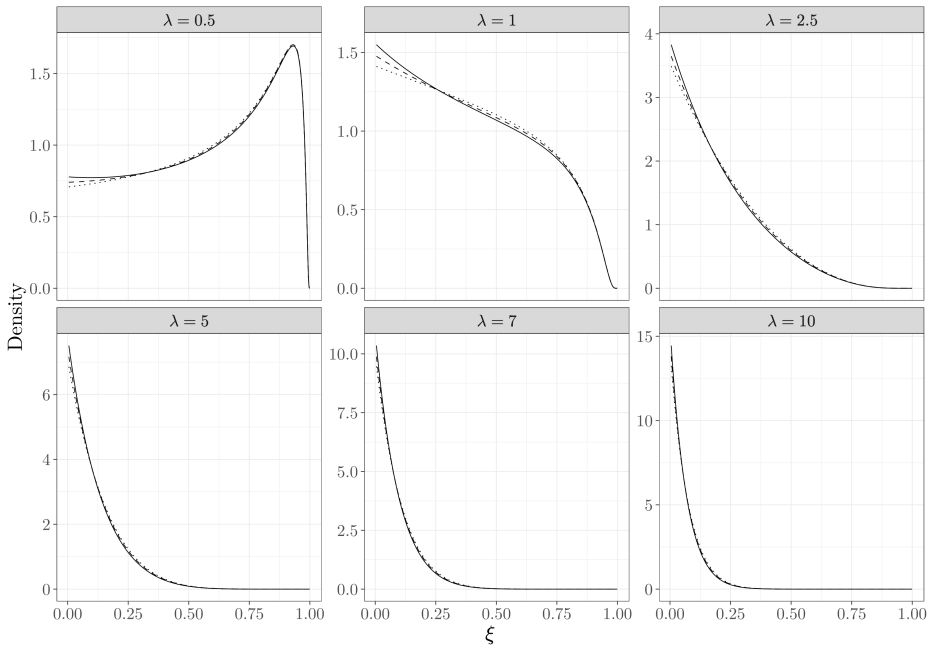
Load INLA library
Create full data frame

Subset data to get exceedances
Create model formula
Define α
Fit the model, gives formula
model likelihoood (GPD)
Give data
Give α
Get summary of the fit

Note: In Practical Session 2 we will learn when it is a good idea to fit a model with this type of linear predictors.

Introduction to INLA – What about the priors?

- In the Bayesian framework, we must define prior distributions over all the model's parameters.
- In the model above, the parameters are ξ, α, β.
- In R-INLA, all the models have default priors that can be modified.
- For example, the default priors for α and β are zero-mean Gaussian distributions with a variance of 10^6 (i.e., a very "flat" prior, reflecting our ignorance about α and β).
- The prior for ξ is a bit more complex and depends on a parameter λ. Figure 2 shows the prior for ξ for different values of λ ($\lambda = 7$ by default).



- In this course (specifically Practical Session 4), we will only use the default priors.

More about INLA

- A requirement to fit models with INLA is that the model belongs to the class of latent Gaussian models.
- But there are other conditions that our model need to meet in order to use INLA.
- We will come back to INLA during Practical 4. There, you would not need to worry about those conditions; they will be met by our model.
- If you want to know more about these conditions, the whole INLA framework and more, visit r-inla.org.

References

1. Castro-Camilo, D., de Carvalho, M., & Wadsworth, J. (2018). Time-varying extreme value dependence with application to leading European stock markets. The Annals of Applied Statistics, 12(1), 283-309.

2. Davison, A. C., Padoan, S. A., & Ribatet, M. (2012). Statistical modeling of spatial extremes. Statistical science, 27(2), 161-186.

3. Smith, R. L. (1989). Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. Statistical Science, 367-377.

4. Genton, M. G., Padoan, S. A., & Sang, H. (2015). Multivariate max-stable spatial processes. Biometrika, 102(1), 215-230.

5. Shooter, R., Ross, E., Ribal, A., Young, I. R., & Jonathan, P. (2022). Multivariate spatial conditional extremes for extreme ocean environments. arXiv preprint arXiv:2201.10451.

6. Smith, R. L. (2003). Statistics of extremes, with applications in environment, insurance and finance. Extreme values in finance, telecommunications and the environment, 1, 78.