EPFL

A Study of Extreme Rainfall in Costa Rica with Multivariate Extreme Value Analysis

Master thesis

Antoine Bourret

Supervisors: Anthony C. Davison, Juan José Leitón-Montero

10.02.2022

Introduction

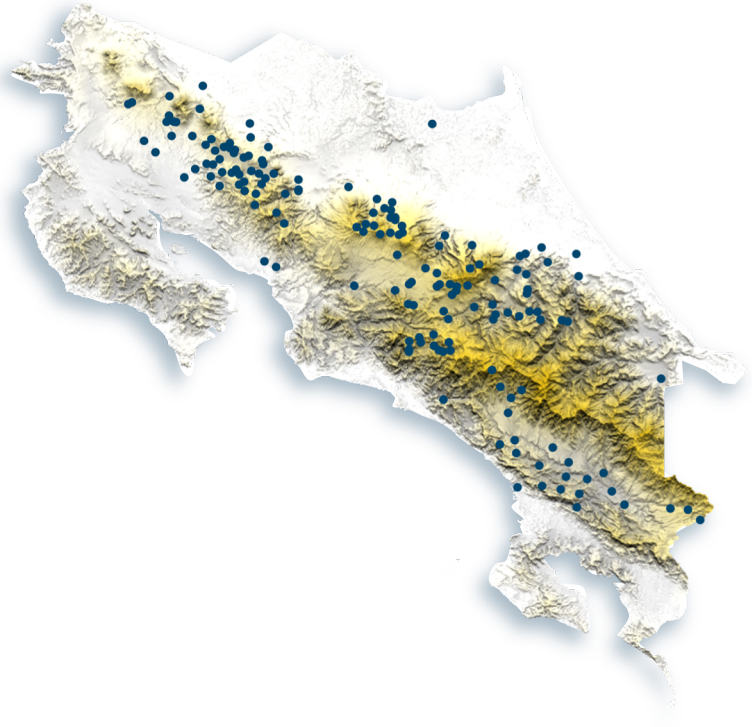
Data

Univariate analysis of extremes

Multivariate analysis of extremes

- Clustering analysis
- Bivariate analysis
- Max-stable processes

Conclusion



Introduction

- Why extreme rainfall?

The study of extreme rainfall events through the construction of **intensity-duration-frequency (IDF)** curves is of great importance for water resource management, drainage basin systems design, and the study of risks of flooding, crop failure and landslide.

- What are intensity-duration-frequency (IDF) curves ?

IDF curves are graphical representations of how often (the frequency) average rainfall intensities (measured in mm/hr) computed over some duration d occur (measured in minutes).

More details can be found in Koutsoyiannis et al. [1998].

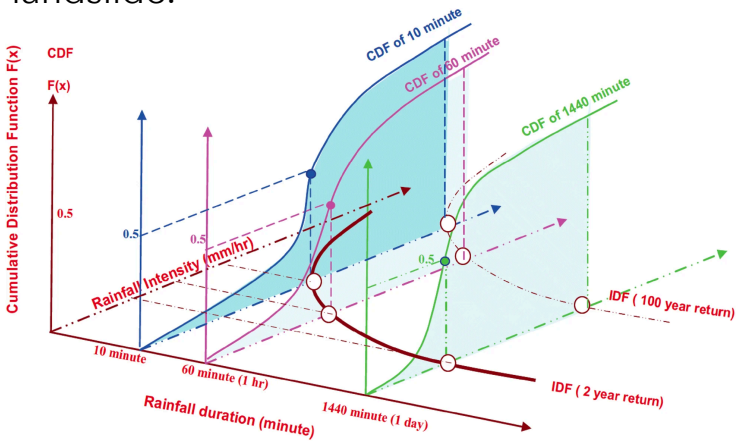
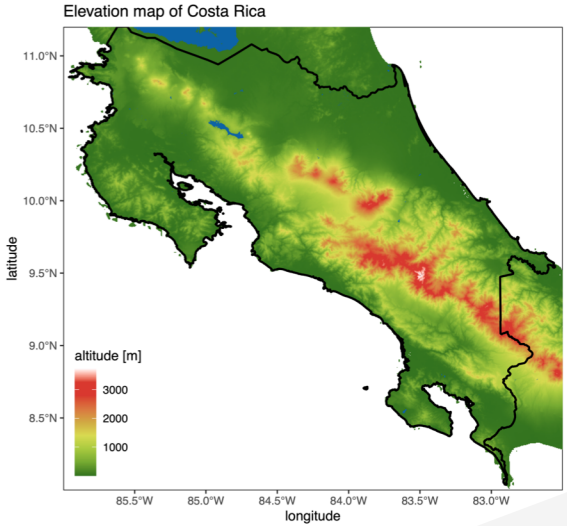
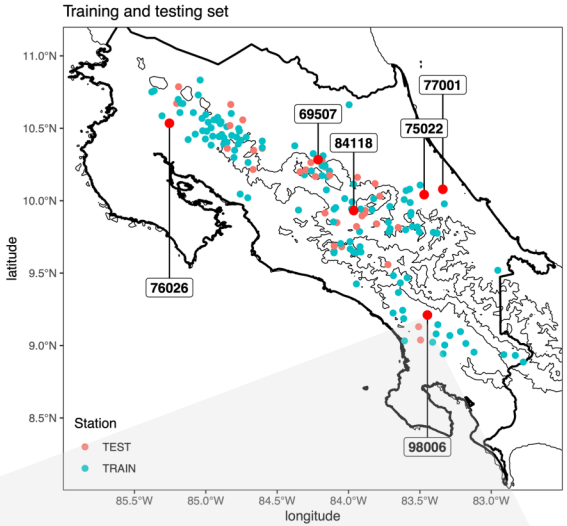


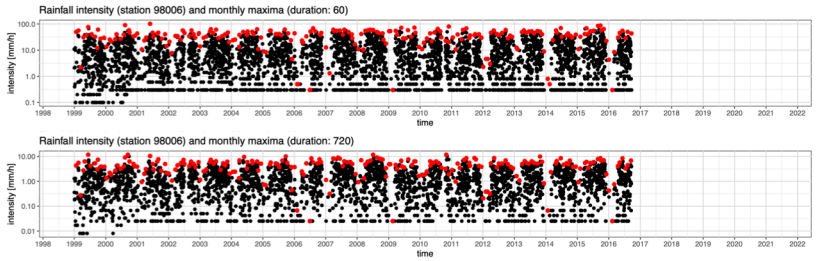
Image from Sun et al. [2019].

- D. Koutsoyiannis, D. Kozonis, and A. Manetas (1998). A mathematical framework for studying rainfall intensity-duration-frequency relationships. In: Journal of Hydrology 206, pp. 118–135.
- Sun, Y., Wendi, D., Kim, D.E. *et al.* (2019). Deriving intensity–duration–frequency (IDF) curves using downscaled in situ rainfall assimilated with remote sensing data. *Geosci. Lett.* 6, 17.

Data

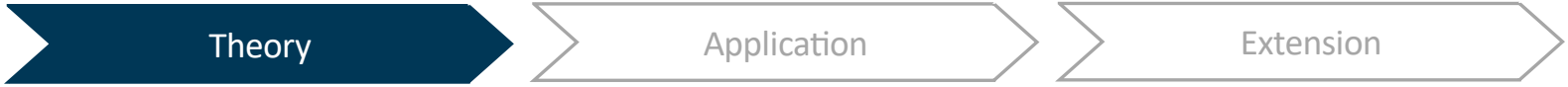






- **daily average rainfall intensities** for 161 stations across Costa Rica, in [mm/hr], provided by the Instituto Costarricense de Electricidad;
- from January 1990 to June 2021, for 10 durations: 5, 10, 15, 30, 60, 120, 180, 360, 720 and 1440 minutes;
- **topographical data**, from The General bathymetric Chart of the Oceans (GEBCO, https://download.gebco.net) and Japan Aerospace Exploration Agency (JAXA, https://global.jaxa.jp/)

Univariate analysis of extremes



- Why extreme value theory?

The construction of IDF curves relies on a frequency analysis in order to derive exceedance probabilities. Although several distributions can be used to model maximum intensities [Stedinger and Foufoula-Georgiou, 1993], the **generalized extreme value distribution** is particularly relevant for modeling the maximum of several random processes, such as mean precipitation.

- Setting
  - Assume that we have access to a sequence $X_1, \dots, X_n \sim F$ of independent daily observations, and that we are interested in modelling the distribution of the maximum $M_n = \max(X_1, \dots, X_n)$.
  - In practice, the distribution F is unknown, and $\mathbb{P}(M_n \leq x) = F^n(x)$ does not provide any useful information as $n \to \infty$.
  - Similar to the central limit theorem, one can consider studying the convergence of the centered and scaled quantities $(M_n - b_n)/a_n$, with $b_n \in \mathbb{R}$ and $a_n \in \mathbb{R}_+$.

- Stedinger, J. and Foufoula-Georgiou, E. (1993). Frequency analysis of extreme events. *Hand- book of Hydrology* 18.

Theorem (Extremal types)
If there exist sequences of constants $b_n \in \mathbb{R}$ and $a_n \in \mathbb{R}_+$ such that

$$\mathbb{P}\left[\frac{(M_n - b_n)}{a_n} \leq x\right] \longrightarrow G(x), \qquad n \longrightarrow \infty,$$

where G is a non-degenerate distribution fonction, then G must be the **generalized extreme value distribution** (GEV), given by

$$G(x) = \begin{cases} \exp\left[-\{1 + \xi(x-\mu)/\sigma\}_+^{-1/\xi}\right], & \xi \neq 0 \\ \exp\left[-\exp\left(-\frac{x-\mu}{\sigma}\right)\right], & \xi = 0 \end{cases}$$

for $x \in \left\{x: 1 + \frac{\xi(x-\mu)}{\sigma} > 0\right\}, \xi, \mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_+$.

▸ Inference

The distribution G has 3 parameters: the location μ, the scale σ and the shape ξ. The latter is crucial to determine the rate of tail decay.

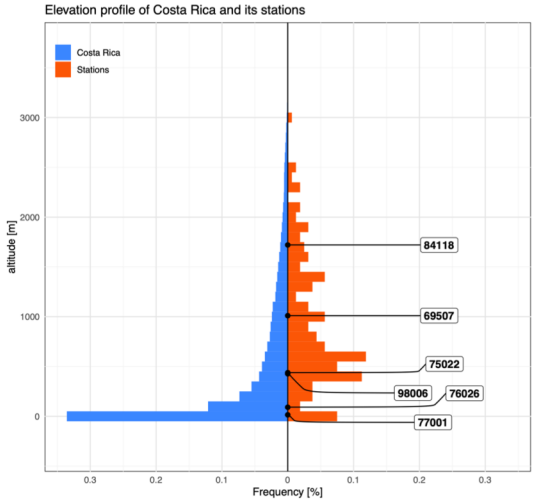
In practice, given a sequence $X_1, X_2, \ldots$ of observations, one can partition them into blocks of equal length, and create the set of maxima $Z_j = \max_{i \in B_j} X_j$ with B_j the set of indices corresponding to the j-th block.

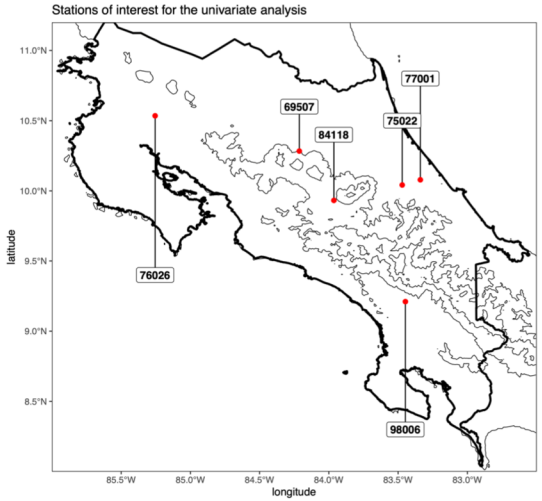
Assuming the independence of the $Z_j, j = 1, \ldots, K$, one can estimate the GEV parameters by maximizing the log-likelihood

$$l(\mu, \sigma, \xi; Z) = -K \log \sigma - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{K} \log\left(1 + \xi \frac{z_i - \mu}{\sigma}\right) - \sum_{i=1}^{K} \left(1 + \xi \frac{z_i - \mu}{\sigma}\right)^{-\frac{1}{\xi}}, \qquad \xi \neq 0.$$

▸ Application

The first step involves the construction of IDF curves for 6 of the 161 precipitation gauges. All these stations started to provide measurements from January 1999, for about 15 to 20 years of daily data.





We modeled monthly maxima for different months and durations. Note that the stations have been treated separately in this case. Because of the **seasonality of monthly maxima** and the effect of the duration on rainfall intensities, the GEV parameters were modeled as

$$\mu_{t,d,s} = f_{s,\mu}^{cc,cs}\,[month(t), d] + \mu_s$$

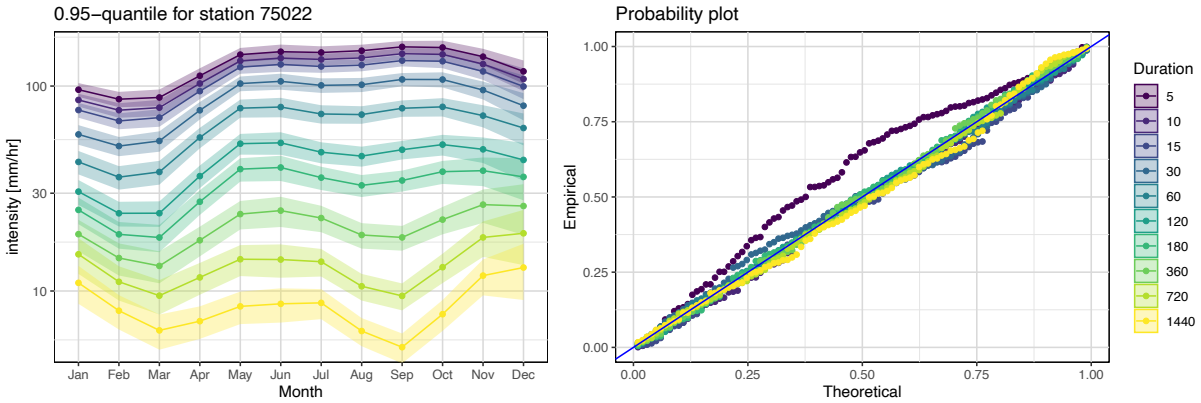
$$\log \sigma_{t,d,s} = f_{s,\sigma}^{cc,cs}\,[month(t), d] + \sigma_s$$

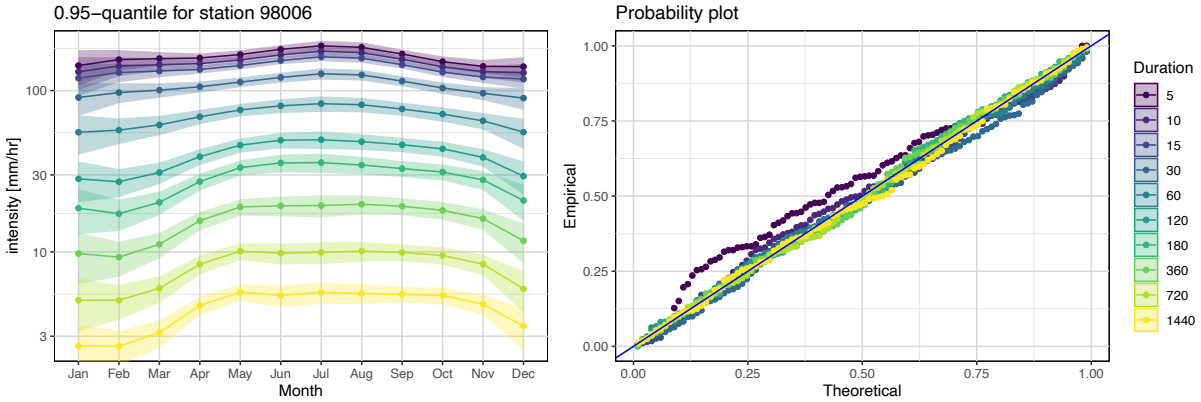
$$\xi_{t,d,s} = f_{1,s,\xi}^{cc}\,[month(t)] + f_{2,s,\xi}^{cs}\,(d) + \xi_s,$$

where $f_{\dots}^{cc}$ and $f_{\dots}^{cs}$ are respectively cyclical cubic and cubic spline functions, and $f_{\dots}^{cc,cs}$ is the tensor product of spline functions.

This model allows for the effect of the duration to change depending on the month, so that rainfall intensity peaks can occur during different month depending on the duration.

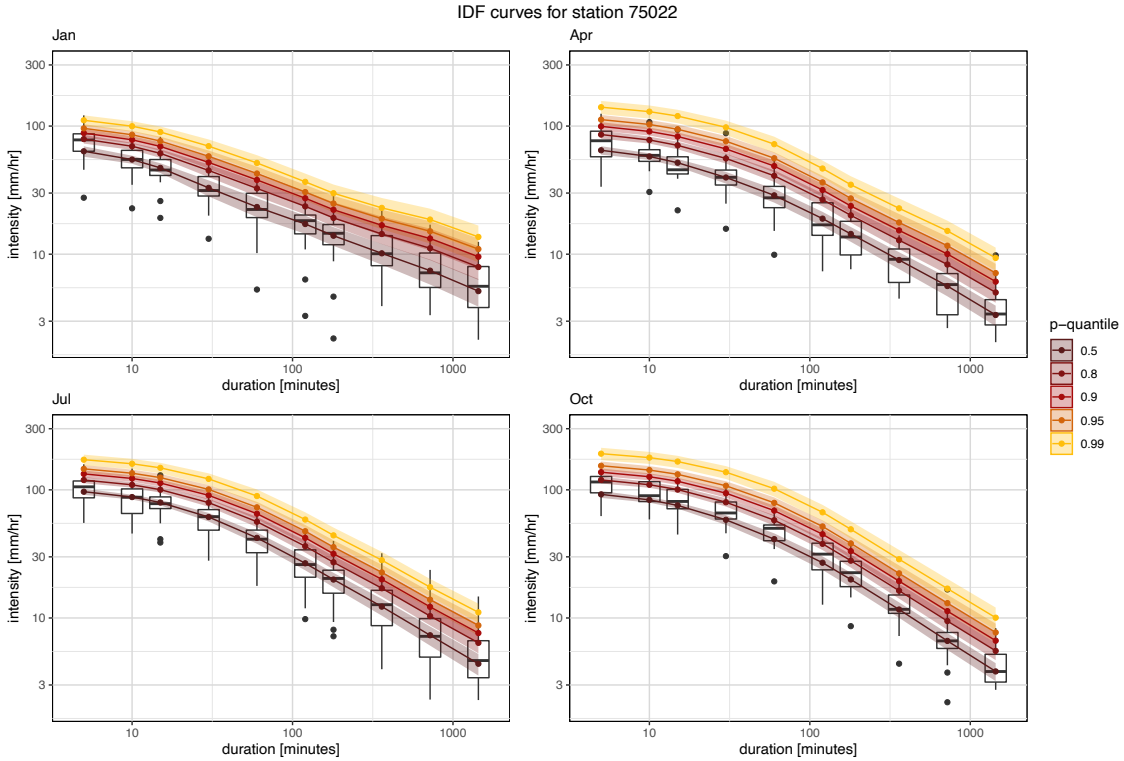
We also used the **deviance statistic** and the **Kolmogorov-Smirnov test** to simplify this model for each station. Overall, the effect of the duration on the shape can be discarded.





- Station 75022 has 20-year return levels with different seasonality depending on the duration (with a peak between June and October for short durations, and between November and January for long ones). This is also observed for stations 69507 and 77001.

- For station 98006, rainfall intensities return levels peak between May and September, for all durations (same for stations 76026 and 84118)



Example of return levels for multiple periods (2, 5, 10, 20 and 100 years), for all durations and different months.

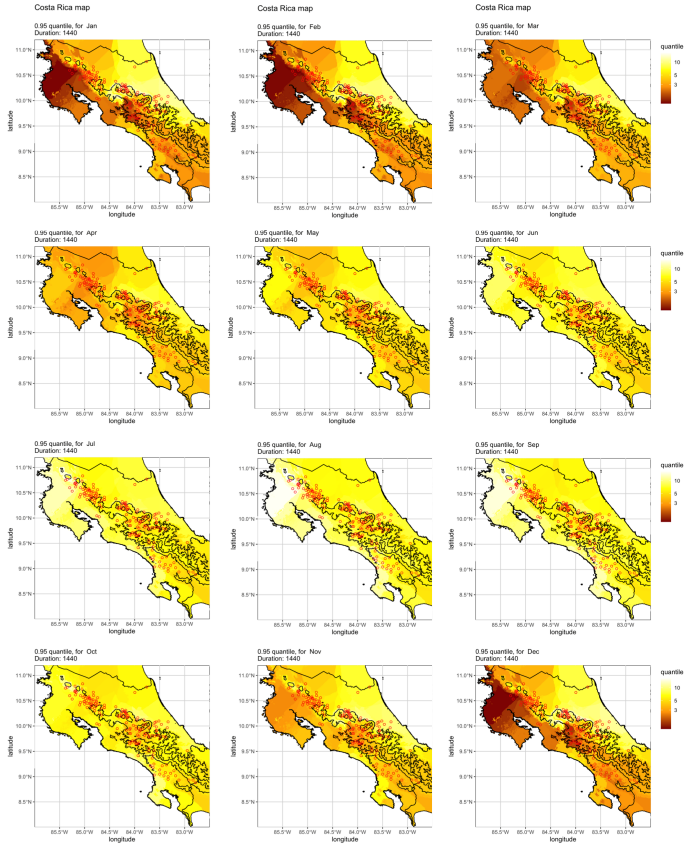
- How to create IDF curves at any location?

So far, we created IDF curves only for locations that can provide observations. But what about locations that are hard to access, or with few precipitation gauge ?

Approach 1: k-NN

The first approach uses the ***k*-nearest neighbours algorithm** to extend GEV parameters at any location.

We fit the models described in the previous section at each station, and then use the standardized longitude, latitude and altitude as covariates to compute the **Euclidian distance**. The parameter $k = 2$ was selected with cross-validation, and confidence bands were obtained with **bootstrapping**.

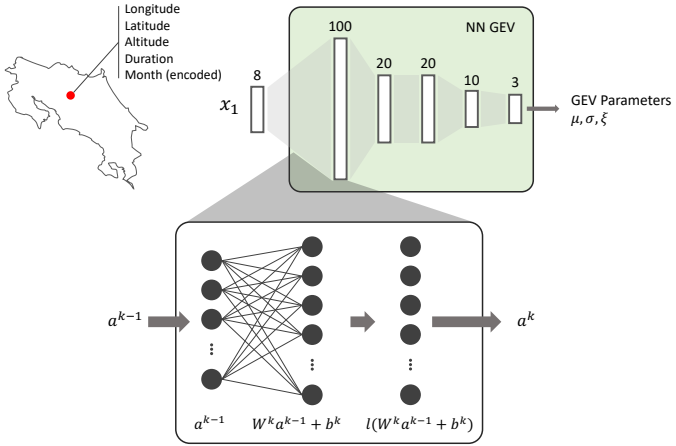


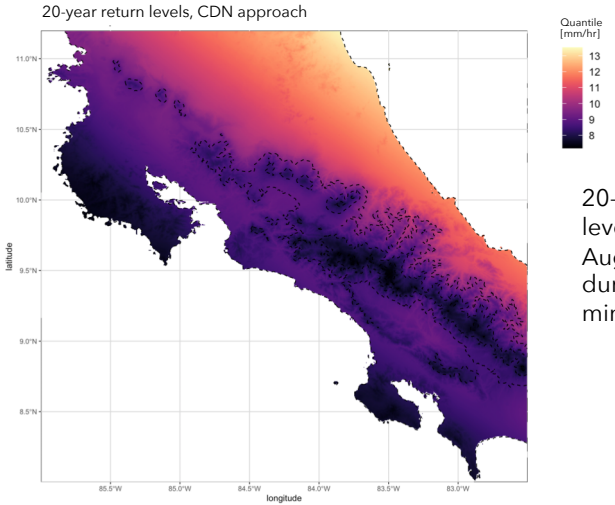
Approach 2: conditional neural network (CDN)

The second approach is based on Cannon [2010].

It uses a **deep neural network** to predict the GEV parameters, taking the longitude, latitude, altitude, duration and month as inputs. The neural network is trained by maximizing the GEV log-likelihood.

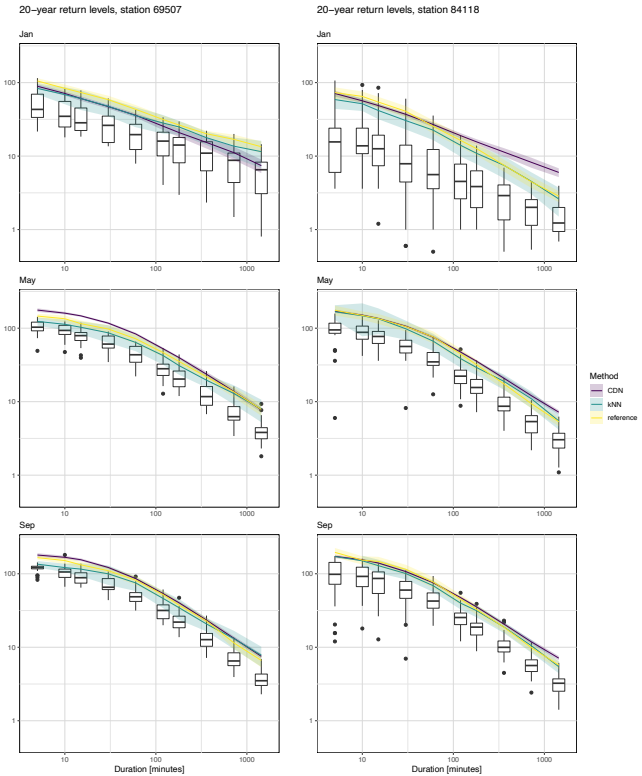
Confidence bands were created with **bootstrapping**. However, this involves the training of multiple models, which can be computationally expensive.





20-year return levels (for August), for a duration of 1440 minutes.

- Cannon, A. J. (2010). A flexible nonlinear modelling framework for nonstationary generalized extreme value analysis in hydroclimatology. *Hydrological Processes* 24: 673–685



- k-NN or CDN approach ?
  - The k-NN approach is faster to train when taking into account the bootstrapping procedure.
  - The CDN approach offers more flexibility regarding the modelling of joint effects of the duration, location and seasonality.

This figure shows 20-year return levels for station 69507 and 84118, for three months. The reference model corresponds to the one fitted for this specific station.

Estimates of the return levels for high durations might differ, especially for January. The k-NN approach has narrower confidence bands.

We also compared the models with a modified version of the **Quantile Skill Index** [Ulrich et al., 2020] that account for the uncertainty of the return levels.

- [1]Ulrich, J., Jurado, O., Peter, M., Scheibel, M. and Rust, H. (2020). Estimating idf curves consistently over durations with spatial covariates. *Water* 12(11): 3119

Conclusion of the univariate approach

- Drawbacks of the univariate analysis:
  - it considers that observation are independent from one station to another, while in practice two very close stations are likely to observe extreme events of the same order of magnitude, during the same periods;
  - one can not use observations of extreme precipitation of a group of stations derive the probability distribution of extreme rainfall of another station. This could be very useful for regions with few precipitation gauges.

In the second part of this project, we studied the relationship of monthly maximum precipitation between multiple sites using two methods: a **bivariate approach** based on the Hüsler-Reiss model, and a modified version of the **non-stationary max-stable** process of Huser and Genton [2016].

Multivariate analysis of spatial extremes

Clustering analysis → Bivariate analysis → Max-stable processes

In order to better understand the relationship and the dependence of two locations in a stochastic process, one can use the variogram. For extremes dependence, as the distributions can be heavy-tailed, it is common to use the **F-madogram**. It is defined as

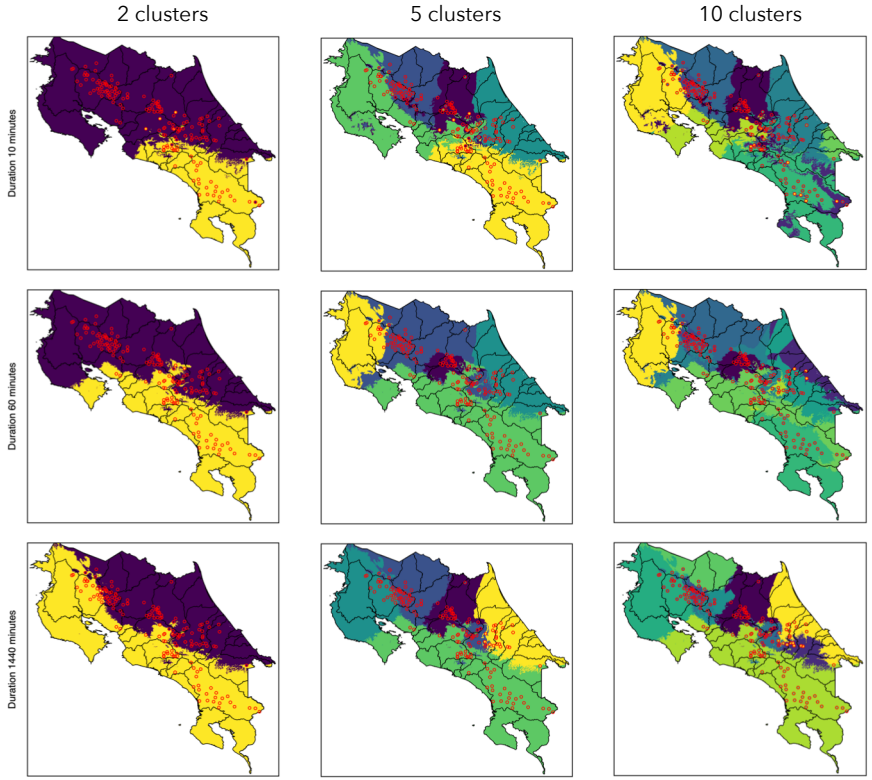
$$v_F(x_1 - x_2) = \mathbb{E}\{|F[Z(x_1)] - F[Z(x_2)]|\}/2$$

where Z is the spatial stochastic process with cumulative distribution function F, and x_1 and x_2 are two locations in the space. A natural estimator is

$$\hat{v}_F(x_1 - x_2) = \frac{1}{2N}\sum_{i=1}^{N}\left|\hat{F}[z_i(x_1)] - \hat{F}[z_i(x_2)]\right|$$

We then used the **hierarchical clustering** approach of Saunders et al. [2019] that is based on the distance $d(x_1, x_2) = \hat{v}_F(x_1 - x_2)$, and the ***k*-nearest neighbours algorithm** to extend the classification to any other location.

- Saunders, K. R., Stephenson, A. G. and Karoly, D. J. (2019). A Regionalisation Approach for Rainfall based on Extremal Dependence. *arXiv: Applications*.



- Results
  - The regions have similar topographical properties, and are not too dislocated across the country.
  - As the duration increases, the Pacific/Caribbean demarcation is more visible.

- Why is this useful?
  - Useful for the implementation of risk anticipation measures for certain areas of the country.
  - This can be used as a preliminary step to defined regions that will then be separately studied using max-stable processes, for example.

Multivariate analysis of spatial extremes

- Setting

Assume that we are given a sequence $(X_1, Y_1), (X_2, Y_2), \ldots$ of independent and identically distributed vectors with distribution function $F_{X,Y}$, and let

$$M_n = \left(\max_{j=1,\ldots,n} X_j \, , \, \max_{j=1,\ldots,n} Y_j \right).$$

We want to study the behaviour of M_n as $n \longrightarrow \infty$.

Theorem

Let (Z_1, Z_2) be the linearly rescaled component-wise maxima of n independent vectors (X_j, Y_j), transformed to have limiting **unit Fréchet marginal distributions**. If

$$\lim_{n\to\infty} \mathbb{P}(Z_1 \leq z_1, Z_2 \leq z_2) = H(z_1, z_2), \qquad z_1, z_2 > 0,$$

where H is a non-degenerate distribution function, then

$$H(z_1, z_2) = \exp[-V(z_1, z_2)], \qquad z_1, z_2 > 0,$$

where the function V is called the **exponent measure**, and we can write

$$V(z_1, z_2) = 2\int_0^1 \max\left(\frac{w}{z_1}, \frac{1-w}{z_2}\right) Q(\mathrm{d}w) = 2\mathbb{E}\left\{\max\left(\frac{w}{z_1}, \frac{1-w}{z_2}\right)\right\},$$

with $W \sim Q$ an **angular distribution** function on $[0,1]$ such that

$$\mathbb{E}(W) = \int_0^1 wQ(\mathrm{d}w) = 1/2.$$

▸ Inference

In practice, one has access to a series $(x_1, y_1), \dots (x_n, y_n)$ of independent vectors, that are used to create the component-wise block maxima $(z_{1,1}, z_{2,1}), \dots (z_{1,m}, z_{2,m})$.

Because the margins are often not distributed according to a **Fréchet distribution**, one needs to transform the variable in order to use the theorem. Suppose that

$$Z_{k,j} = \mathrm{GEV}(\mu_{k,j}, \sigma_{k,j}, \zeta_{k,j}), \qquad j = 1, \dots, m, \qquad k = 1,2.$$

The transformed variables

$$Z^*_{k,j} = \left[1 + \xi_{k,j} \frac{Z_{k,j} - \mu_{k,j}}{\sigma_{k,j}}\right]^{1/\xi_{k,j}}$$

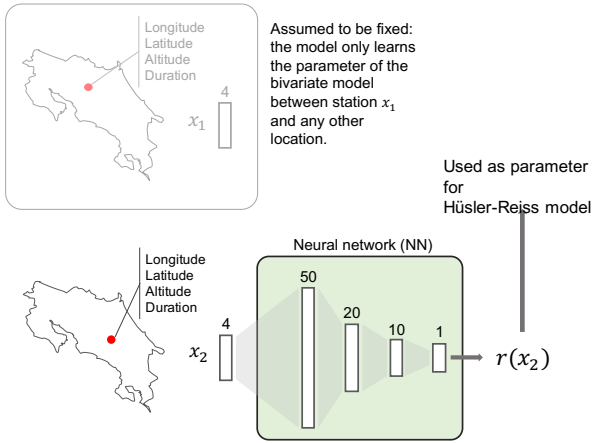
are then distributed according to the standard Fréchet distribution. The bivariate model is then fitted using **maximum likelihood**, with a choice of dependence function V.

Extending the bivariate model

We extended the Hüsler–Reiss model with a conditional neural network. The distribution is

$$\mathbb{P}(Z(x_1) \leq z_1, Z(x_2) \leq z_2) = \exp\left\{-\frac{1}{z_1}\Phi\left(\frac{1}{r}+\frac{r}{2}\log\frac{z_2}{z_1}\right)-\frac{1}{z_2}\Phi\left(\frac{1}{r}+\frac{r}{2}\log\frac{z_1}{z_2}\right)\right\}, \qquad r > 0,$$

and the parameter r is approximated with a deep neural network that takes as input the longitude, latitude, altitude and duration of transformed rainfall at two stations. The first station x_1 is fixed and only the second station x_2 changes during the training.

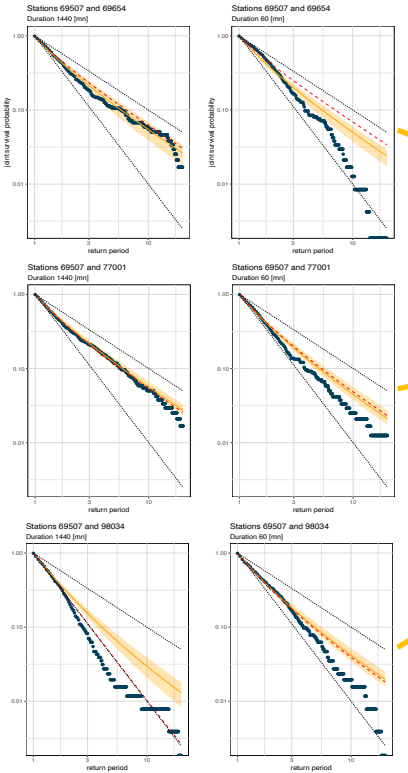


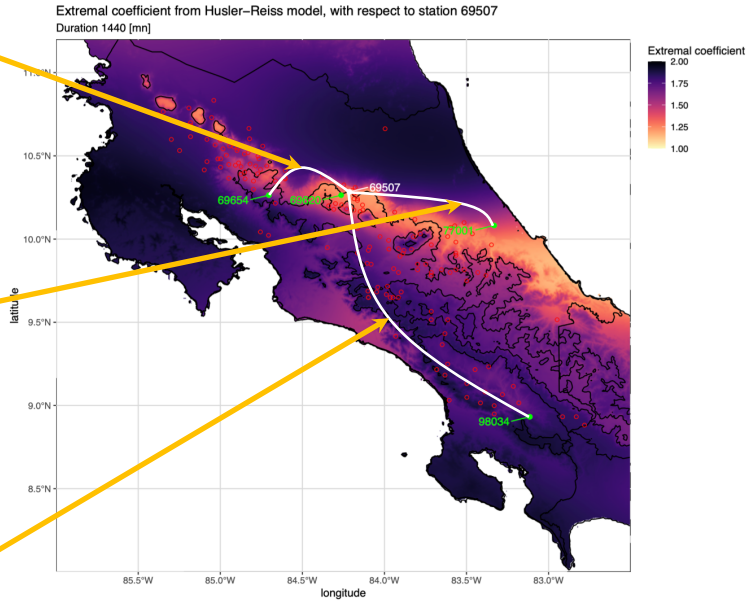
- Results

The **joint survival probability**
$\mathbb{P}(Z(x_1) > z, Z(x_2) > z)$
can be used to assess if extreme events at two locations are likely to occur at the same time.

For most stations, the model provided a reasonable approximation for the joint survival probability, though it struggles for the independent case.

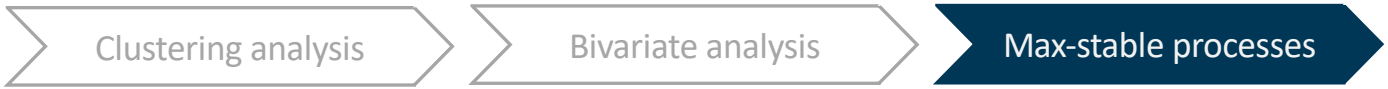
Joint survival probability





extremal coefficient: $\theta = V(1,1)$, so that
$\mathbb{P}(Z_1 \leq z, Z_2 \leq z) = \exp\big(-\mathrm{V}(\mathrm{z},\mathrm{z})\big)$
$= \exp(-1/\mathrm{z})^{V(1,1)}, \qquad z > 0.$

Multivariate analysis of spatial extremes



Even if the bivariate approach allows to quantify to which extent some high precipitation levels observed at a particular station can also be observed at any other location in Costa Rica, it nevertheless involves to fit one model per precipitation gauge.

An extension to allow for any two location was undertaken, but did not lead to any useful results. This motivates the use of **max-stable processes**, which offer a better understanding of the behaviour of extremal dependence across Costa Rica.

Definition

A stochastic process $\{Z(x), x \in \mathcal{X}\}$ is said to be **max-stable** if there exist sequences $A_{N_x} > 0, B_{N_x}$ for $N > 0$ and $x \in \mathcal{X}$, such that if $Z^{(1)}(x), \ldots, Z^{(N)}(x)$ are N independent copies of the process, and if one defines $\{Z^*(x), x \in \mathcal{X}\}$ as

$$Z^*(x) = \frac{\max_{1 \le i \le N} Z^{(i)}(x) - B_{N_x}}{A_{N_x}}, \qquad x \in \mathcal{X}.$$

Then $\{Z^*(x), x \in \mathcal{X}\}$ is identical in distribution to $\{Z(x), x \in \mathcal{X}\}$.

There exist multiple classes of max-stable processes, as highlighted by Smith [1990]. One can in particular use a **spectral representation**.

- Smith, R. L. (1990). Max-stable processes and spatial extremes (Unpublished manuscript)

In the following, we will focus on the **extremal *t* max-stable process** [Nikoloulopoulos et al., 2009] with spectral representation

$$Z^*(x) = \max_i [\varepsilon_i c_v \max\{0, Y_i(x)\}^v],$$

with

$$c_v = 2^{1-v/2}\pi^{1/2}\left[\Gamma\left\{\frac{(v+1)}{2}\right\}\right]^{-1},$$

$\{\varepsilon_i, i \geq 1\}$ realisations of a Poisson process on $\mathbb{R}_+^*$ with intensity measure $\varepsilon^{-2}\mathrm{d}\varepsilon$, and $v > 0$. $Y_i(x)$ are realisations of a standard Gaussian process with correlation function $\rho(x_1, x_2)$ and Γ is the gamma function.

The model is fitted by maximizing the **pairwise log-likelihood**

$$l_p(\mathbf{z}, \theta) = \sum_{k=1}^{N} \sum_{(i,j)\in P_k} \log f(z_{k,i}, z_{k,j};\ \theta) \qquad (1)$$

where θ are the parameters of the model, $\boldsymbol{z}$ is the set of all observations, P_k is the set of pairs of locations (i, j) for which the process was observed with realizations $(z_{k,i}, z_{k,j})$, for $k = 1, \ldots, N$, and f is the joint bivariate density.

- Nikoloulopoulos, A., Joe, H. and Li, H. (2009). Extreme value properties of multivariate t copulas. *Extremes* 12: 129–148

- The bivariate distribution of the **extremal *t* max-stable process** is

$$\mathbb{P}(Z(x_1) \leq z_1, Z(x_2) \leq z_2) = \exp\left\{-\frac{1}{z_1}T_{v+1}\left[r\left(\frac{z_2}{z_1}\right)\right] - \frac{1}{z_2}T_{v+1}\left[r\left(\frac{z_1}{z_2}\right)\right]\right\},$$

where T_v is the Student t cumulative distribution function with v degrees of freedom, and

$$r(t) = r_{x_1,x_2}(t) = \frac{t^{1/v} - \rho(x_1, x_2)}{(v+1)^{-1/2}[1-\rho(x_1,x_2)]^{1/2}}.$$

- In order to account the presence of multiple **local effects** and **weather systems**, Huser and Genton [2016] proposed the **non-stationary dependence structure**

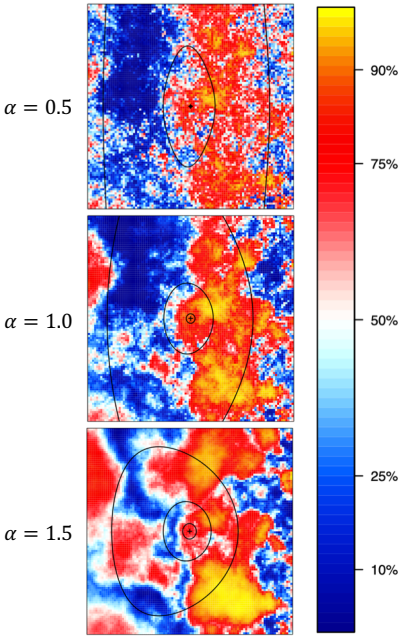
$$\rho(x_1, x_2) = |\Omega_{x_1}|^{\frac{1}{4}}|\Omega_{x_2}|^{\frac{1}{4}}\left|\frac{\Omega_{x_1} + \Omega_{x_2}}{2}\right|^{-\frac{1}{2}} R\left(Q^{\frac{1}{2}}_{x_1,x_2}\right),$$

and

$$R(h) = \exp(-h^{\alpha}), \qquad h > 0, \alpha \in]0,2].$$

The parameter α controls the roughness of the random field, and the quadratic form Q is defined as

$$Q_{x_1,x_2} = (x_1 - x_2)^T\left(\frac{\Omega_{x_1} + \Omega_{x_2}}{2}\right)^{-1}(x_1 - x_2).$$



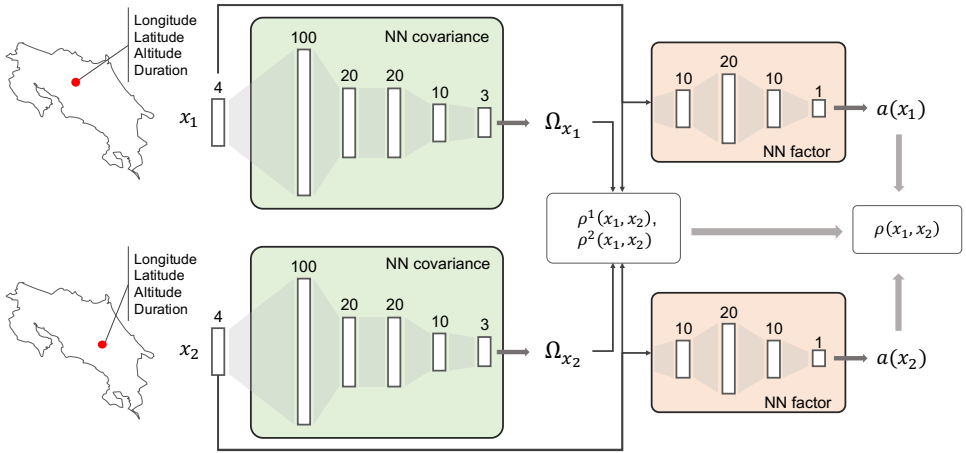
Example of random field with different roughness parameter. Images taken from Huser and Genton [2016].

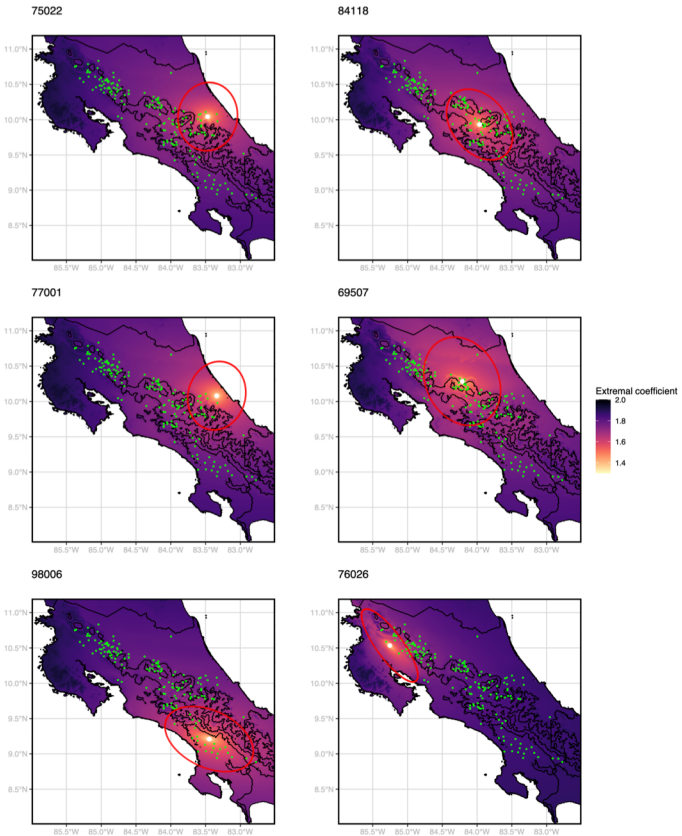
- Huser, R. and Genton, M. (2016). Non-Stationary Dependence Structures for Spatial Ex- tremes. *Journal of Agricultural, Biological, and Environmental Statistics* 21: 470–491.

As suggested by Huser and Genton [2016], we used a **mixture of max-stable processes** with two different roughness parameters α, and correlation function

$$\rho(x_1, x_2) = \frac{a(x_1)a(x_2)\rho^1(x_1, x_2) + \{1 - a(x_1)\}\{1 - a(x_2)\}\rho^2(x_1, x_2)}{\sqrt{[a(x_1)^2 + \{1 - a(x_1)\}^2][a(x_2)^2 + \{1 - a(x_2)\}^2]}}.$$

The covariance matrix Ω_x at each location x is approximated by a deep neural network, as well as the function $a(x)$. The neural network is trained by maximizing the pairwise log likelihood (1).





The max-stable approach offers several advantages compared to the bivariate one:

- the fit of a unique model that can provide a measure of dependence between any two locations;
- a more intuitive understanding of dependence between extreme events, following the rainfall-storm interpretation of Smith [1990];
- the possibility to directly generate simulations from the max-stable process, which can be used to estimate the joint dependence of extrema between more than two locations.

- Smith, R. L. (1990). Max-stable processes and spatial extremes (Unpublished manuscript)

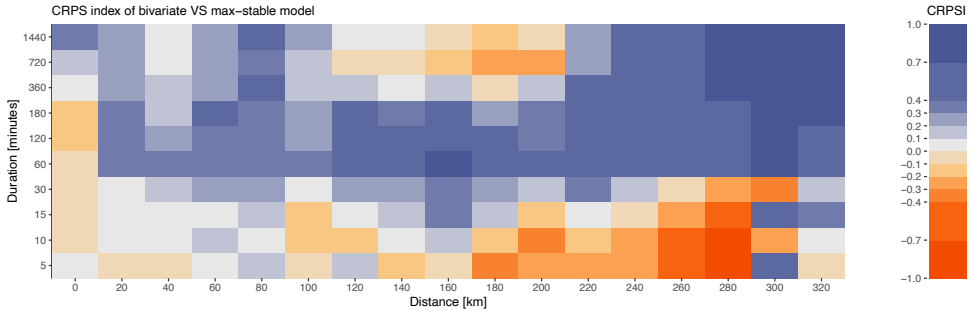
- Bivariate or max-stable approach ?

We compared the bivariate and max-stable models using the **continuous rank probability score index (CRPSI)**, which is inspired by the CRPS [Gneiting and Raftery, 2007] and the QSI used in the univariate analysis.

Given an observed value x and a distribution function F, the CRPS is defined as

$$CRPS(F,x) = \int_{-\infty}^{\infty} \left[F(y) - 1_{\{y \geq x\}}\right]^2 \mathrm{d}y\,.$$

For both bivariate and max-stable approaches, the CRPS was computed based on the empirical estimates of the extremal coefficient.



Positive values of CRPS index (in blue) indicate improvements of the bivariate model compared to the non-stationary max-stable process, while negative ones (red) an improvement of the max-stable to the bivariate model.

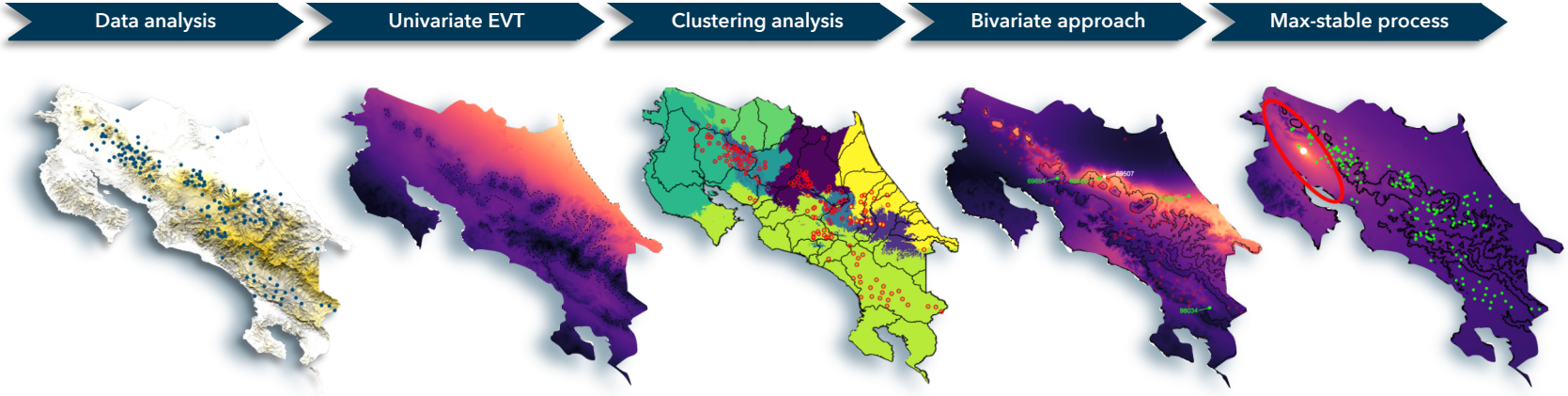
- Gneiting, T. and Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* 102: 359–378

Conclusion

Conclusion

- In this project, extreme rainfall events were characterized through **intensity-duration-frequency** curves that were constructed with **extreme value theory**. We fitted the generalized extreme value distribution with non-stationary parameters for monthly rainfall maxima for each precipitation gauge separately.
- In order to produce estimates of return levels for different periods at any location in Costa Rica, we used two methods (the $\boldsymbol{k}$**-NN** and **CDN** approaches), and compared them with the Quantile Skill Index.
- We finally studied the **relationship of extreme rainfall** between different locations, starting from a **regionalization approach** to indentify regions with similar behaviour, and then by quantifying the extremal dependence with a **bivariate model** and a **non-stationary max-stable process**.

Conclusion



▸ This study highlights regions with **strong seasonality** and a **clear demarcation between the Pacific and Caribbean coasts**, especially for extreme precipitation events that extend over long periods. The use of **deep neural networks** allowed to model the local effects and the different weather systems present in Costa Rica, both from an univariate and multivariate point of view.

Thank you !

References

- D. Koutsoyiannis, D. Kozonis, and A. Manetas (1998). A mathematical framework for studying rainfall intensity-duration-frequency relationships. In: Journal of Hydrology 206, pp. 118–135.
- Sun, Y., Wendi, D., Kim, D.E. *et al*. (2019). Deriving intensity–duration–frequency (IDF) curves using downscaled in situ rainfall assimilated with remote sensing data. *Geosci. Lett.* 6, 17.
- Stedinger, J. and Foufoula-Georgiou, E. (1993). Frequency analysis of extreme events. *Hand- book of Hydrology* 18.
- Cannon, A. J. (2010). A flexible nonlinear modelling framework for nonstationary generalized extreme value analysis in hydroclimatology. *Hydrological Processes* 24: 673–685.
- Ulrich, J., Jurado, O., Peter, M., Scheibel, M. and Rust, H. (2020). Estimating idf curves consistently over durations with spatial covariates. *Water* 12(11): 3119.
- Saunders, K. R., Stephenson, A. G. and Karoly, D. J. (2019). A Regionalisation Approach for Rainfall based on Extremal Dependence. *arXiv: Applications*.
- Nikoloulopoulos, A., Joe, H. and Li, H. (2009). Extreme value properties of multivariate t copulas. *Extremes* 12: 129–148.
- Huser, R. and Genton, M. (2016). Non-Stationary Dependence Structures for Spatial Ex- tremes. *Journal of Agricultural, Biological, and Environmental Statistics* 21: 470–491.
- Smith, R. L. (1990). Max-stable processes and spatial extremes (Unpublished manuscript).
- Gneiting, T. and Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* 102: 359–378.