Building and simulating complex geostatistical models for climate and environmental sciences

Denis Allard,
[with Lionel Benoit, **Lucia Clarotto**, Xavier Emery,
Grégoire Mariethoz, Saïd Obakrim, Thomas Opitz]

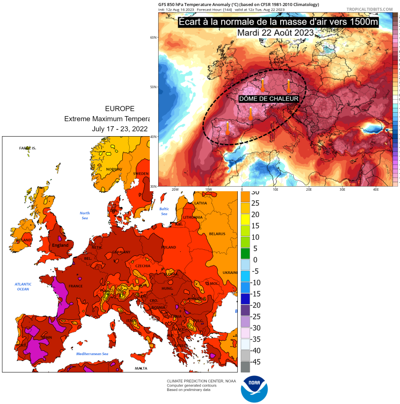
Biostatistique et processus Spatiaux (BioSP), INRAE
Avignon, France

Math for Mathematics for planet Earth (M4E) Workshop
11-12 of November, 2024

Increasing frequency of "natural" disasters

Storm Boris 14th of Septembre 2024

Increasing frequency of ”natural” disasters



Heat Waves June 2022 and August 2023

Non extreme climate events with huge impact

2016: Up to 30% yield loss on wheat in the "Breadbasket"

Compound events

Zscheischler et al., 2020

”A combination of multiple drivers and/or hazards that contributes to societal or environmental risk ”

- Pre-conditionned
- Multivariate
- Temporal succession
- Spatially distributed

Stochastic simulation as a tool to address this complexity

Outline of the talk

1. Some reminders on spatial statistics
2. A multivariate, spatio-temporal Stochastic Weather Generator
3. A stochastic generator for heat waves

Statistical model

Often: trend + GP + noise

$$Z(\mathbf{s},t) = \mu(\mathbf{s},t) + Y(\mathbf{s},t) + \epsilon(\mathbf{s},t), \quad (\mathbf{s},t) \in D \times T \subset \mathbb{R}^d \times \mathbb{R}$$

- For example:

$$h[\mu(\mathbf{s},t)] = \sum_{k=1}^{p} \beta_k f_k(X_k(\mathbf{s})) + \sum_{l=1}^{q} \alpha_l g_l(X_l(t))$$

- $Y(\mathbf{s},t)$ is a centered, second order, stationary Gaussian Process

$$\mathbf{Cov}\{Y(\mathbf{s},t), Y(\mathbf{s}+\mathbf{h}, t+u)\} = C(\mathbf{h}, u)$$

- $\epsilon(\mathbf{s},t)$ random noise with mean 0

Need for **valid and relevant covariance functions**

Positive definiteness

Valid = positive definite

A function $\boldsymbol{C} : \mathbb{R}^d \times \mathbb{R} \mapsto \mathbb{R}^p$ is a matrix-valued stationary covariance function iif $\boldsymbol{C}$ is a **positive definite** matrix-valued function : $\forall n, \forall \mathbf{s}_1, \ldots, \mathbf{s}_n \in \mathbb{R}^d, \ \forall t_1, \ldots, t_n \in \mathbb{R}$ et $\forall \boldsymbol{a}_1, \ldots, \boldsymbol{a}_n \in \mathbb{R}^p$

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{p}\sum_{l=1}^{p} a_{i,k} C_{kl}(\mathbf{s}_j - \mathbf{s}_i, t_j - t_i) a_{j,l} \geq 0$$

- Use covariance functions from known classes, e.g. Matérn

Univariate modeling: Gneiting class

Definition (Gneiting, 2002)

$$C(\mathbf{h}, u) = \frac{1}{(1+\gamma(u))^{\tau}} C_{\infty}\left(\mathbf{h}/(1+\gamma(u))^{b/2}\right)$$

- b is a separability parameter, with $0 \leq b \leq 1$
- γ is an unbounded variogram, e.g.: $\gamma(u) = (u.r_T)^{\alpha}$, with $0 \leq \alpha \leq 2$
- C_{∞} is a covariance function on $\mathbb{R}^d$, $\forall d \geq 1$, e.g. a Matérn covariance function

$$\mathcal{M}(\mathbf{h}; r_S, \nu) = \frac{\sigma^2 2^{1-\nu}}{\Gamma(\nu)} \left(\|\mathbf{h}\|.r_S\right)^{\nu} \mathcal{K}_{\nu}\left(\|\mathbf{h}\|.r_S\right),$$

where $r_S > 0$ is a scale parameter and $\mathcal{K}_{\nu}$ is the modified second order Bessel function of order ν

Multivariate generalization

Theorem (Allard, D., Clarotto, L. and Emery, X., 2022)

$$C_{ij}(\mathbf{h}, u) = \frac{\tau_{ij}}{(\eta_{\mathbf{ij}}(u) + b_{ij}^2)^{\tau}} \mathcal{M}\left(\mathbf{h}; \frac{a_{ij}}{\left(\eta_{\mathbf{ij}}(\mathbf{u}) + b_{ij}^2\right)^{b/2}}, \nu_{ij}\right),$$

- $[C_{ij}(\mathbf{h}, u)]_{i,j=1}^{p}$ is a multivariate space-time covariance under some conditions on the $p \times p$ matrices $\boldsymbol{b}$, $\boldsymbol{a}$, $\boldsymbol{\nu}$ et $\boldsymbol{\tau}$
- $\boldsymbol{\eta}(\mathbf{u})$ is a $p \times p$ matrix-valued unbounded pseudo-variogram on $\mathbb{R}$

- Each variable has its own set of parameters in space and in time
- Illustrated later

SWGs

Adapted from Yiou (2024)

SWG are tools that generate random series of meteorological variables such as precipitation, temperature, wind speed, etc., with statistics similar to those of recorded data:

- Mean, variance, quantiles, skewness, extremes
- Covariance (dependence) between variables
- Temporal dependence / coherence (persistence)
- Spatial dependence / coherence

- Calibrated on recorded series
- Computational efficiency $\Rightarrow$ long series and/or large number of realizations

For what purpose?

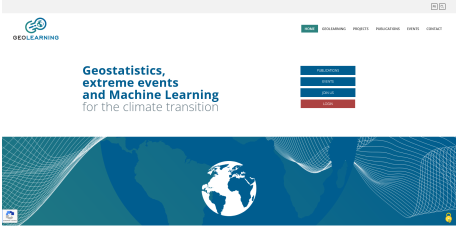
Used in impact studies

Outputs of SWGs are used as inputs in process-based models, e.g. energy demand models, crop models, hydrological models, insurance models, ...

- Assessing complex, non linear, responses to climate in agro-ecological systems
- Explore unmeasured climates
- Explore plant / ecosystem models as functions of climate variability
- Optimal decision under uncertainty: simulate up to year $t + k$, optimize decision
- Disaggregating (downscaling) meteorological variables from GCM outputs

Some challenges that SWGs pose to spatial statistics

- Building models quantifying spatial, temporal and spatio-temporal variations
- Doing stochastic simulations, both for the bulk and for the tail
- Building models and methods for multivariate, spatio-temporal extreme events
- Devising new approaches for assessing return levels of impactful compound events



Our context; the project BEYOND

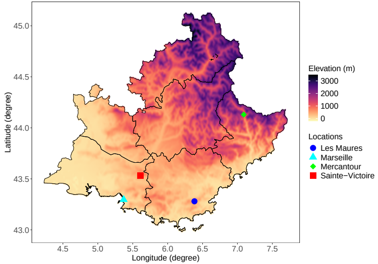
- The **BEYOND** project: towards new tools for epidemiological surveillance (for plants)
- *Xylella Fastidiosa* (Xf) is a plant pathogen propagated by insects
- Major damages: 54,000 ha of dead or uprooted olive orchards in Italy
- Seen in Corsica, Baleares, Tuscany
- Propagation depends on the whole seasonal cycle: not just one ”extreme” event

$\hookrightarrow$ Need for a **stochastic tool able to generate complete cycles over a region**

INRAE département MathNum Biostatistique BIO/Π & Processus Spatiaux GEOLEARNING CHAIRE /// Data Science for the Environment

Our domain of interest

- Region of interest: PACA, highly non-stationary
- 6 daily variables: precipitation, humidity, radiation, wind, min and max temperature
- SAFRAN reanalysis data from 2012 to 2021



General architecture

Two main assumptions

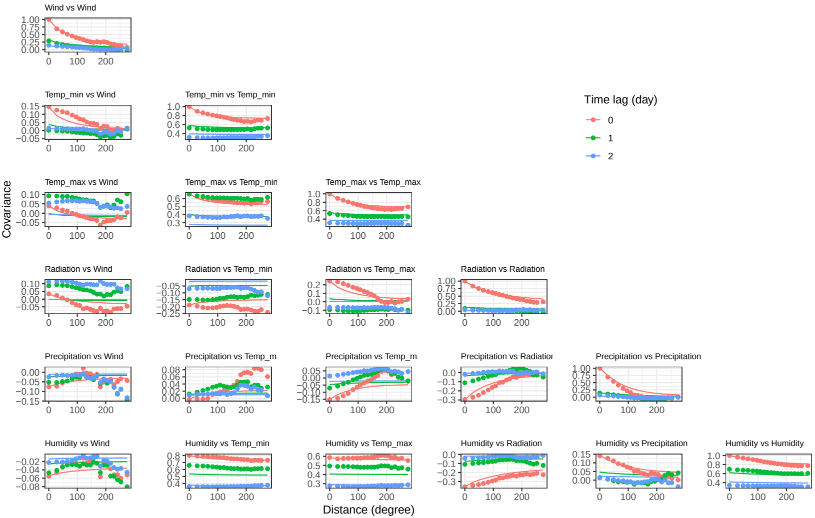
- Finite number of weather types over the region, $k = 1, \ldots, K$
- In each weather type k, the weather variables are modeled as transformed latent Gaussian random fields

$$Y_i(\mathbf{s}, t) = \psi_{k,i,\mathbf{s}}\left(Z_{k,i}(\mathbf{s}, t)\right), \qquad k = X(t)$$

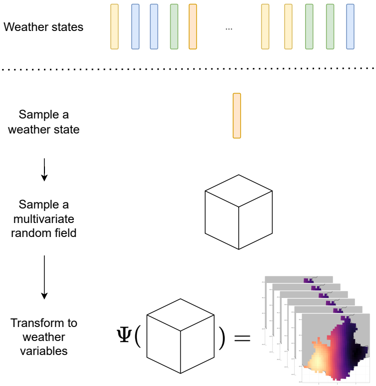
Weather types | Transformation function | Latent Gaussian field

- Weather states are modeled as 1st order Markov chain with transition matrices $\pi(t)$
- We use empirical transformation for $\psi_{k,i}$ with tail adjustments (Peterson and Cavanaugh, 2019)
- Multivariate spatio-temporal GPs for $\boldsymbol{Z} = (Z_i)_{i=1,p}$

Multivariate covariance

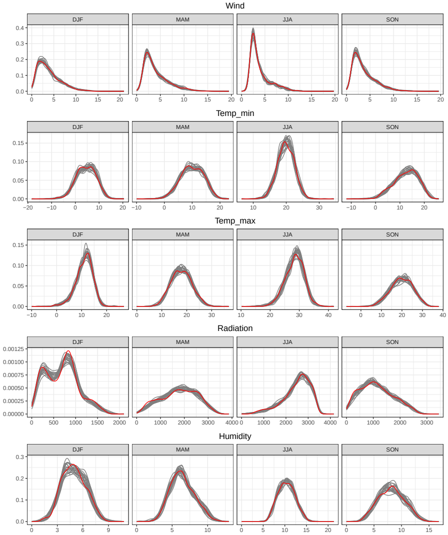


Simulation


Weather states
...
Sample a weather state
Sample a multivariate random field
Transform to weather variables
$\Psi(\ ) =$

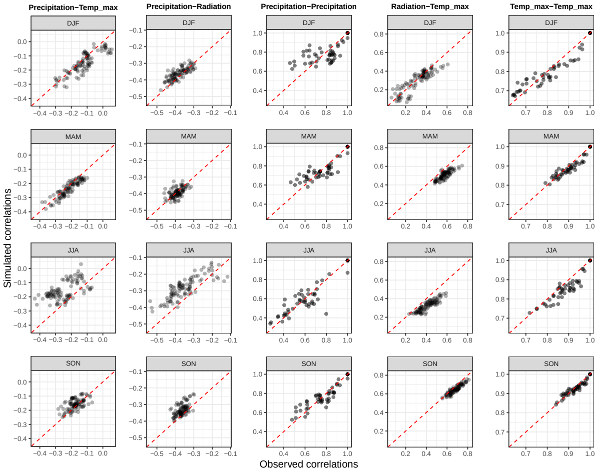
Marginals

All seasons, all continuous variables



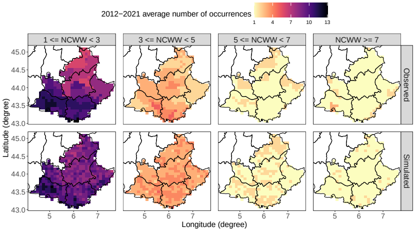
Correlations

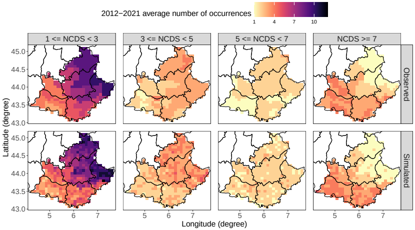
Winter, 10 random locations



Wet and dry spells

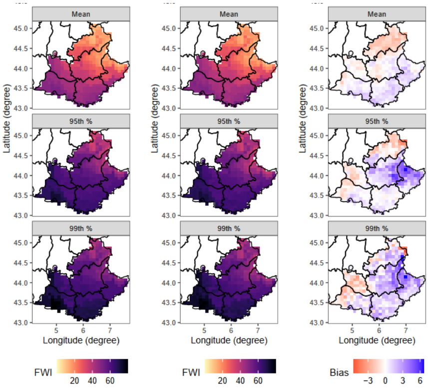
Wet winter spells, Dry summer spells





Fire Weather Index

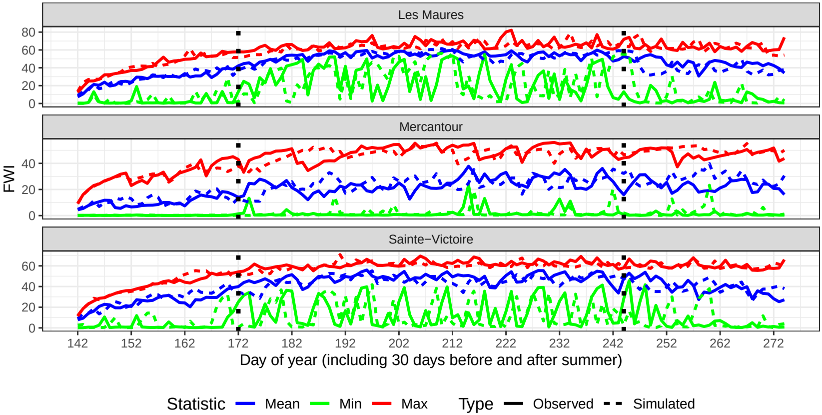
Summer



Obs. Sim.

Fire Weather Index

Summer



Take home messages

- It is possible to design a spatial, multivariate SWG with quite good statistical performances
- One among many possible statistical model

MSTWeatherGen

- Obakrim S., Benoit L., Allard D., Rey J. (2024). MSTWeatherGen: Multivariate Space-Time Weather Generator. R package
  `https://sobakrim.github.io/MSTWeatherGen/index.html`
- Obakrim, S., Benoit, L., & Allard, D. (2024). A multivariate and space-time stochastic weather generator using a latent Gaussian framework.
  `https://hal.science/hal-04715860/`

- Future work: non-stationary covariance; increased persistence, **long period of droughts**
- PhD student (A. Doizé) with P. Naveau and O. Wintenberger

Typology of SWGs

Model based (parametric)

- (+) identification through a set of parameters ⇒ sensitivity analysis
- (+) can create non recorded situations and simulate more extreme conditions than those observed
- (−) temporal and spatial coherence sometimes difficult to reproduce

Resampling / analogs (non parametric)

- (+) compatibility between climatic variables is guaranteed
- (+) statistical features and spatial/temporal are reproduced by construction
- (−) cannot create unobserved meteorological situations
- (−) Implicit assumption: the most extreme observation has been observed

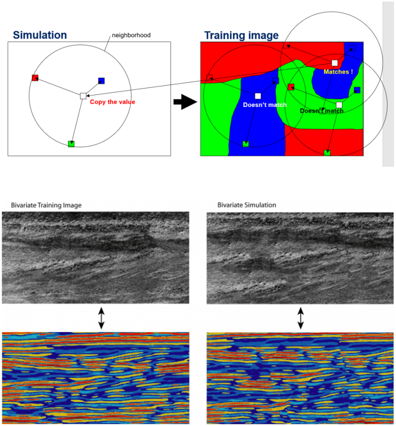
Our approach

- Spatial resampling (Direct Sampling) for the spatial patterns
- Extreme value theory for the extrapolation of very high values

Opitz, T., Allard, D., & Mariethoz, G. (2021). Semi-parametric resampling with extremes. *Spatial Statistics*, 42, 100445.

Spatial resampling

Direct Sampling (DS) (Mariethoz and Caers, 2014)



- Cannot generate values beyond those observed
- Tend to under-represent the extremal dependence

↪ Use extreme value theory as a complement to DS

Pareto processes

Generalized Pareto Distribution

If $X(\mathbf{s}) \sim F_{\mathbf{s}}, \quad \mathbf{s} \in \mathcal{D}$, then

$$X(\mathbf{s}) - u \mid X(\mathbf{s}) > u \sim GPD_{\sigma(\mathbf{s}),\xi(\mathbf{s})} \quad \text{as} \quad u \to \infty$$

Pareto processes (Dombry et Ribatet, 2015)

- Suppose uniform margins for $X^U(\mathbf{s})$
- Consider a homogeneous risk functional $r(X^U)$

$$(1-u)^{-1} X^U(\mathbf{s}) \mid [\mathbf{r}(\mathbf{X^U}) > u] \to Y(\mathbf{s}), \quad \mathbf{s} \in \mathcal{D}, \quad \text{as } u \to 1$$

where

$$Y(\mathbf{s}) = \eta(\mathbf{s}).\mathbf{r}(\mathbf{Y})$$

with

$$\eta(\mathbf{s}) := \frac{Y(\mathbf{s})}{\mathbf{r}(\mathbf{Y})} \perp \mathbf{r}(\mathbf{Y})$$

"Uplift" extremal fields

- We have T independent copies of $X_t(\mathbf{s})$, $t = 1, \ldots, T$
- Consider only realizations such that

$$\mathbf{r}\left(\mathbf{X^U}\right) > u \qquad u \in (0, 1)$$

UpliftExtremeFields

1. Compute $\eta(\mathbf{s}) = X^U(\mathbf{s})/\mathbf{r}\left(\mathbf{X^U}\right)$.
2. Draw $q \sim \mathrm{Unif}(u, 1)$ or set q for a given return period
3. Generate the uplifted field

$$\tilde{X}^U(\mathbf{s}) = q\eta(\mathbf{s}), \quad \mathbf{s} \in \mathcal{D}$$

$\hookrightarrow$ will be used the enrich the dataset with synthetic events more extremes than those observed

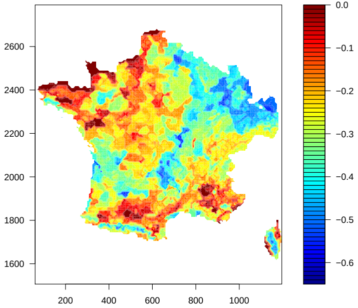
Case study: heat waves in France

Motivation

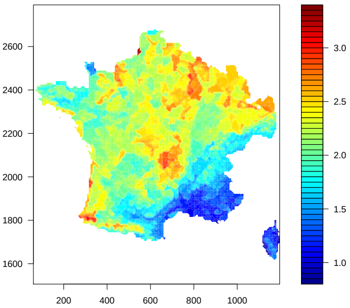
- Absolute breaking temperature record in France, on 28th of June, 2019: **45.9°C** at Gallargues, Gard (previous record was 44.1°C)

- SAFRAN reanalysis data for T_{max}, from 2010 to 2016, June-September
- In each mesh, standardization to uniform
- $r = \mathbf{med}(X_t^U)$

GPD parameters

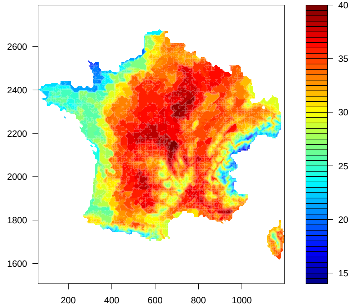


Shape

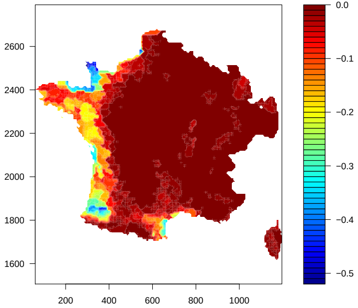


Scale

Most extreme event



Original



Uniform

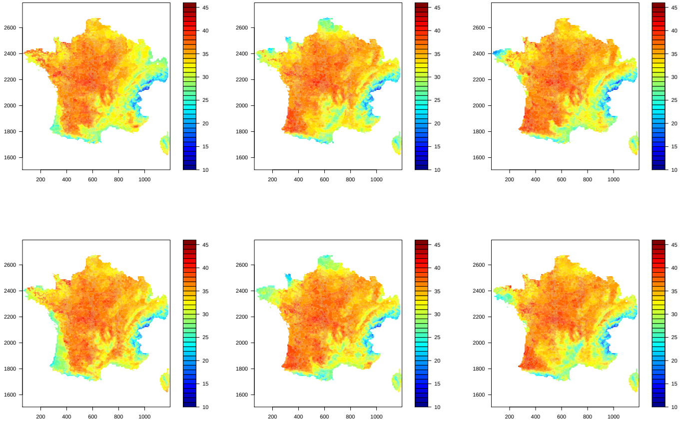
Generating new heatwaves

Resample first, back-transform second

1. Generate uplifted extreme episodes (on the uniform scale)
2. Perform DS on uniform scales to create new spatial patterns
3. Back-transform on Temp scale using $F_{\mathbf{s}}^{-1}$

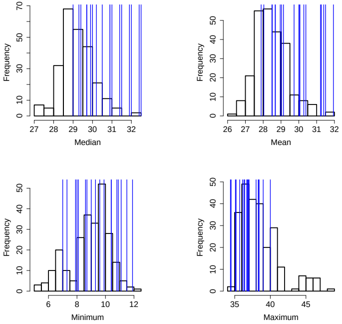
- Non stationarity is properly accounted for
- q is chosen according to a 10 year return period

New realizations



Can we validate?

- Functional risk is r = median on U
- We selected the 30 most extreme events (2011-2016) wrt to r
- The 20 highest are kept aside for validation
- The 10 lowest are used for training
- We use $u = 0.92$, which is the lower bound of the validation set; return period is 17 days during summer
- 250 simulations are generated using DS



Some final words

Heatwaves

- One of the few approaches combining non-parametric and parametric methods on extremes

Current work on SWGs within the Chair Geolearning

- Working on long period of rainfalls and droughts
- Simulation of precipitations using generative approaches
- Simulation of extreme flows on a river system
- Coming up with an open library