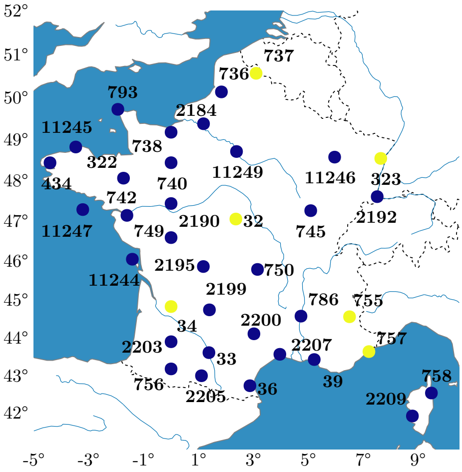
Spatio-temporal generation of precipitations using a spatially correlated Bernoulli and hidden Markov model

Caroline Cognot (EDF & MIA-PS)
Supervision : Liliane Bel (MIA-PS), Sylvie Parey (EDF)
David Métivier (INRAE - MISTEA)

AgroParisTech Talents d'une planète soutenable

Contents

Multisite Rainfall Weather Generators



S = 37 stations, ECA&D

Objective: Build a stochastic weather generator

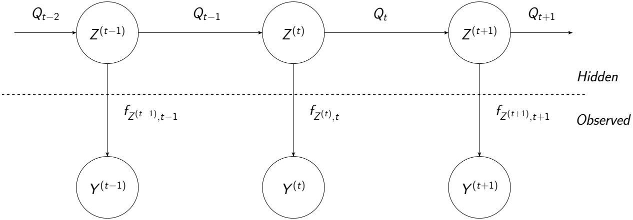
- For the multisite rain occurrence $Y^{(t)} = (Y_1^{(t)}, \ldots, Y_S^{(t)})$
- Then for precipitation intensity $R^{(t)} = (R_1^{(t)}, \ldots, R_S^{(t)})$

→ Reproduce the spatial-temporal structure of the data.

- In particular **large scales dry/wet episodes**

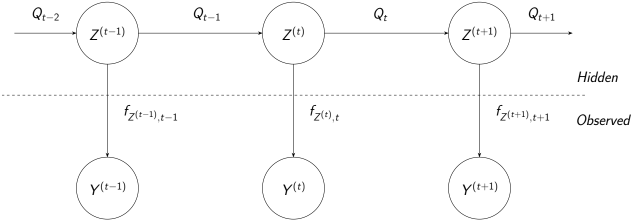
Multisite HMM based model

Rain occurrence $Y^{(t)} = (Y_1^{(t)}, \dots, Y_S^{(t)}) \in \{0,1\}^S$ - Unobserved weather type $Z^{(t)} \in \{1, \dots, K\}$



Multisite HMM based model

Rain occurrence $Y^{(t)} = (Y_1^{(t)}, \ldots, Y_S^{(t)}) \in \{0,1\}^S$ - Unobserved weather type $Z^{(t)} \in \{1, \ldots, K\}$



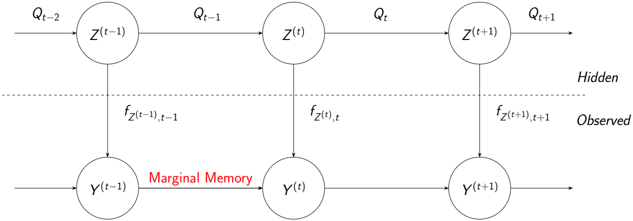
Conditional independence (Zucchini et al. 1991, Gobet et al. 2024)

$$\mathbb{P}\left(Y = y \mid Z = z^{(t)}\right) = f_{z^{(t)},t}(y) = \prod_{s=1}^{S} \left(y_s \lambda_{Z^{(t)},t,s} + (1 - y_s)(1 - \lambda_{Z^{(t)},t,s})\right)$$

With $\lambda_{k,t,i} = \mathbb{P}(Y_i^{(t)} = 1 | Z^{(t)} = k)$ for $i \in 1, \ldots, \mathcal{S}$

Multisite HMM based model

Rain occurrence $Y^{(t)} = (Y_1^{(t)}, \dots, Y_S^{(t)}) \in \{0,1\}^S$ - Unobserved weather type $Z^{(t)} \in \{1, \dots, K\}$



Conditional independence (Zucchini et al. 1991, Gobet et al. 2024)

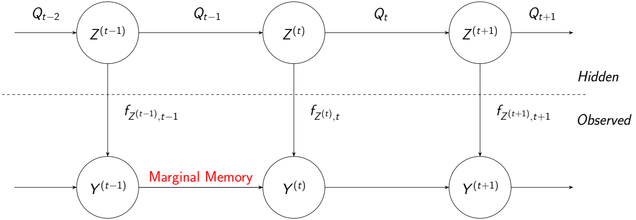
$$\mathbb{P}\left(Y = y \mid Z = z^{(t)}, Y^{(t-1)} = y^{t-1}\right) = f_{z^{(t)},t}(y) =$$

$$\prod_{s=1}^{S}\left(y_s \lambda_{Z^{(t)},t,s,y_s^{t-1}} + (1-y_s)(1-\lambda_{Z^{(t)},t,s,y_s^{t-1}})\right)$$

With $\lambda_{k,t,i,y_i^{t-1}} = P(Y_i^{(t)} = 1 | Z^{(t)} = k, Y_i^{(t-1)} = y_i^{t-1})$ for $i \in 1, \dots, \mathcal{S}$

Multisite HMM based model

Rain occurrence $Y^{(t)} = (Y_1^{(t)}, \ldots, Y_S^{(t)}) \in \{0,1\}^S$ - Unobserved weather type $Z^{(t)} \in \{1, \ldots, K\}$

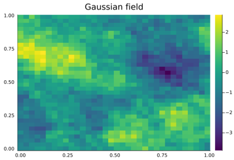


Conditional independence (Zucchini et al. 1991, Gobet et al. 2024)

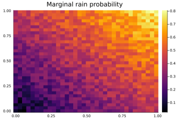
– Stations must be "far apart enough": 10 stations in the paper

\+ Correlations between stations are captured by the weather types Z

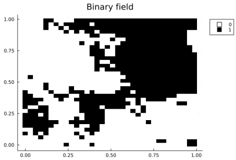
How to generate binary correlated variables



$X \sim \mathcal{N}(0, C_\theta)$: latent

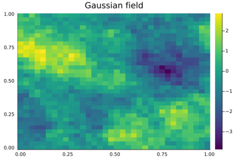


$\xrightarrow{\lambda_s}$

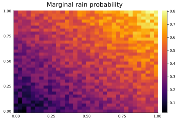


$$\forall s, Y_s = \begin{cases} 1 & \text{if } X_s \leq \Phi^{-1}(\lambda_s) \\ 0 & \text{else} \end{cases}$$

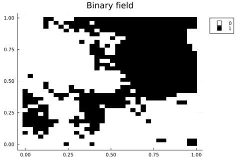
How to generate binary correlated variables



$X \sim \mathcal{N}(0, C_\theta)$: latent



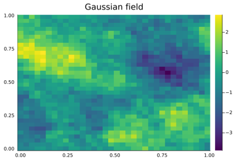
$\xrightarrow{\lambda_s}$



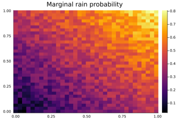
$$\forall s, Y_s = \begin{cases} 1 & \text{if } X_s \leq \Phi^{-1}(\lambda_s) \\ 0 & \text{else} \end{cases}$$

Parameters : $\lambda_i = \mathbb{P}(Y_i = 1)$ for $i \in 1, \ldots, \mathcal{S}$ and θ the parameters of the latent covariance.

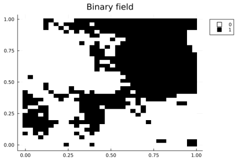
How to generate binary correlated variables



$X \sim \mathcal{N}(0, C_\theta)$: latent



$\xrightarrow{\lambda_s}$



$$\forall s,\ Y_s = \begin{cases} 1 & \text{if } X_s \leq \Phi^{-1}(\lambda_s) \\ 0 & \text{else} \end{cases}$$

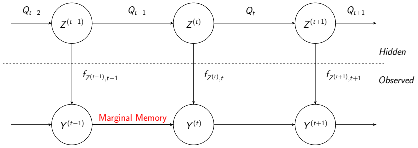
Parameters : $\lambda_i = \mathbb{P}(Y_i = 1)$ for $i \in 1, \ldots, \mathcal{S}$ and θ the parameters of the latent covariance.

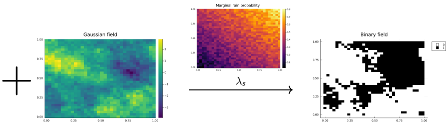
Associated distribution

$$f(y, \theta) = \int_{a_1}^{b_1} \cdots \int_{a_\mathcal{S}}^{b_\mathcal{S}} f_X((x_1, ..., x_\mathcal{S}))d(x_1, ..., x_\mathcal{S})$$

With $a_i = -\infty$ if $Y_i = 1, \Phi^{-1}(\lambda_i)$ else, $b_i = \infty$ if $Y_i = 0, \Phi^{-1}(\lambda_i)$ else

Multisite model





$X \sim \mathcal{N}(0, C_\theta)$: latent

$$\forall s, Y_s = \begin{cases} 1 & \text{if } X_s \leq \Phi^{-1}(\lambda_s) \\ 0 & \text{else} \end{cases}$$

- \+ No restriction on the stations distance
- ± Correlations between stations are captured by the weather types Z and C_θ
- \- The emission is more complicated !

HMM : Maximum likelihood estimation

- Hidden states $\rightarrow$ Expectation-Maximization (EM) algorithm
- Seasonal parameters $\theta(t)$

!! High dimensional integrals

HMM : Maximum likelihood estimation

- Hidden states $\rightarrow$ Expectation-Maximization (EM) algorithm
- Seasonal parameters $\theta(t)$

!! High dimensional integrals

$\rightarrow$ Tricks make the problem computable

Example: Maximize composite pairwise likelihood $\sum_{\text{pairs}i,j} w_{ij} \log L((y_i, y_j); \theta)$ instead of full likelihood $\log L((y_1, \dots, y_{\mathcal{S}}); \theta)$ during the **M** step of the EM algorithm.

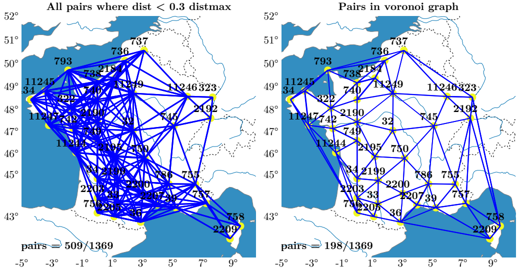
HMM : Maximum likelihood estimation

- Hidden states $\rightarrow$ Expectation-Maximization (EM) algorithm
- Seasonal parameters $\theta(t)$

!! High dimensional integrals

$\rightarrow$ Tricks make the problem computable

Example: Maximize composite pairwise likelihood $\sum_{\text{pairs}i,j} w_{ij} \log L((y_i, y_j); \theta)$ instead of full likelihood $\log L((y_1, \ldots, y_S); \theta)$ during the **M** step of the EM algorithm.



Adding the Rain intensity : marginals

For now : we get $Z^{(t)}$ the state, $(Y_1^{(t)}, ..., Y_{\mathcal{S}}^{(t)})$ the occurence of rain.

What about the amount of rain $(R_1^{(t)}, ..., R_{\mathcal{S}}^{(t)})$?

Adding the Rain intensity : marginals

For now : we get $Z^{(t)}$ the state, $(Y_1^{(t)}, ..., Y_{\mathcal{S}}^{(t)})$ the occurence of rain.

What about the amount of rain $(R_1^{(t)}, ..., R_{\mathcal{S}}^{(t)})$?

Generalized Pareto Distribution

GPD models exceedances over a threshold u:

$$H(r;\sigma,\xi) = \begin{cases} \frac{1}{\sigma}\left(1+\frac{\xi r}{\sigma}\right)^{-1/\xi-1}, & \xi \neq 0, \\ \frac{1}{\sigma} e^{-r/\sigma}, & \xi = 0 \end{cases}$$

Adding the Rain intensity : marginals

For now : we get $Z^{(t)}$ the state, $(Y_1^{(t)}, ..., Y_{\mathcal{S}}^{(t)})$ the occurence of rain.

What about the amount of rain $(R_1^{(t)}, ..., R_{\mathcal{S}}^{(t)})$?

Generalized Pareto Distribution

GPD models exceedances over a **large** threshold u:

$$H(r;\sigma,\xi) = \begin{cases} \frac{1}{\sigma}\left(1+\frac{\xi r}{\sigma}\right)^{-1/\xi-1}, & \xi \neq 0, \\ \frac{1}{\sigma} e^{-r/\sigma}, & \xi = 0 \end{cases}$$

- Drawback: needs a **large** threshold u
- Bulk of the distribution must be modeled **separately**

Adding the Rain intensity : marginals

For now : we get $Z^{(t)}$ the state, $(Y_1^{(t)}, ..., Y_{\mathcal{S}}^{(t)})$ the occurence of rain.

What about the amount of rain $(R_1^{(t)}, ..., R_{\mathcal{S}}^{(t)})$?

Generalized Pareto Distribution

GPD models exceedances over a **large** threshold u:

$$H(r;\sigma,\xi) = \begin{cases} \frac{1}{\sigma}\left(1+\frac{\xi r}{\sigma}\right)^{-1/\xi-1}, & \xi \neq 0, \\ \frac{1}{\sigma} e^{-r/\sigma}, & \xi = 0 \end{cases}$$

Extended Generalized Pareto Distribution, Naveau et al. (2016),Gamet et al. (2022)

EGPD models exceedances over a **low** threshold u:

$$K(r;\sigma,\xi) = G(H(r;\sigma,\xi))$$

Adding the Rain intensity : marginals

For now : we get $Z^{(t)}$ the state, $(Y_1^{(t)}, \ldots, Y_{\mathcal{S}}^{(t)})$ the occurence of rain.

What about the amount of rain $(R_1^{(t)}, \ldots, R_{\mathcal{S}}^{(t)})$?

Generalized Pareto Distribution

GPD models exceedances over a **large** threshold u:

$$H(r;\sigma,\xi) = \begin{cases} \frac{1}{\sigma}\left(1+\frac{\xi r}{\sigma}\right)^{-1/\xi-1}, & \xi \neq 0, \\ \frac{1}{\sigma} e^{-r/\sigma}, & \xi = 0 \end{cases}$$

Extended Generalized Pareto Distribution, Naveau et al. (2016),Gamet et al. (2022)

EGPD models exceedances over a **low** threshold u:

$$K(r;\sigma,\xi) = G(H(r;\sigma,\xi))$$

- Need appropriate $G : [0,1] \to [0,1]$: Truncated Beta of Gamet et al. (2022)
- Only need to model the very low values separately

Adding the Rain Intensity: Spatio-temporal Link

Occurrence probability and truncation bounds.
Let

$$\Phi(T(s,n)) = 1 - \lambda^{(n)}_{z^{(n)},s,y_s^{(n-1)}} = \mathbb{P}\Big(Y_s^{(n)} = 1 \mid Z^{(n)} = z^{(n)},\ Y_s^{(n-1)} = y_s^{(n-1)}\Big),$$

which defines the rain/no–rain threshold $T(s,n)$.

$$\begin{aligned} Y_s^{(n)} = 0: &\qquad \ell(s,n) = -\infty, \quad u(s,n) = T(s,n), \\ Y_s^{(n)} = 1: &\qquad \ell(s,n) = T(s,n), \quad u(s,n) = \infty. \end{aligned}$$

Adding the Rain Intensity: Spatio-temporal Link

Occurrence probability and truncation bounds.
Let

$$\Phi(T(s,n)) = 1 - \lambda^{(n)}_{z^{(n)},s,y_s^{(n-1)}} = \mathbb{P}\Big(Y_s^{(n)} = 1 \mid Z^{(n)} = z^{(n)},\ Y_s^{(n-1)} = y_s^{(n-1)}\Big),$$

which defines the rain/no–rain threshold $T(s,n)$.

$$\begin{aligned} Y_s^{(n)} = 0: &\qquad \ell(s,n) = -\infty, \quad u(s,n) = T(s,n), \\ Y_s^{(n)} = 1: &\qquad \ell(s,n) = T(s,n), \quad u(s,n) = \infty. \end{aligned}$$

Latent truncated Gaussian field :

$$X_R(\cdot,n) \sim TMVN(0,\ C_R(\cdot),\ \boldsymbol{\ell}, \boldsymbol{u}),$$

$=$ $X_R(s,n)$ is marginally standard Gaussian.

Adding the Rain Intensity: Spatio-temporal Link

Occurrence probability and truncation bounds.
Let

$$\Phi(T(s,n)) = 1 - \lambda^{(n)}_{z^{(n)},s,y_s^{(n-1)}} = \mathbb{P}\Big(Y_s^{(n)} = 1 \mid Z^{(n)} = z^{(n)},\ Y_s^{(n-1)} = y_s^{(n-1)}\Big),$$

which defines the rain/no–rain threshold $T(s,n)$.

$$\begin{aligned} Y_s^{(n)} = 0: &\qquad \ell(s,n) = -\infty, \quad u(s,n) = T(s,n), \\ Y_s^{(n)} = 1: &\qquad \ell(s,n) = T(s,n), \quad u(s,n) = \infty. \end{aligned}$$

Latent truncated Gaussian field :

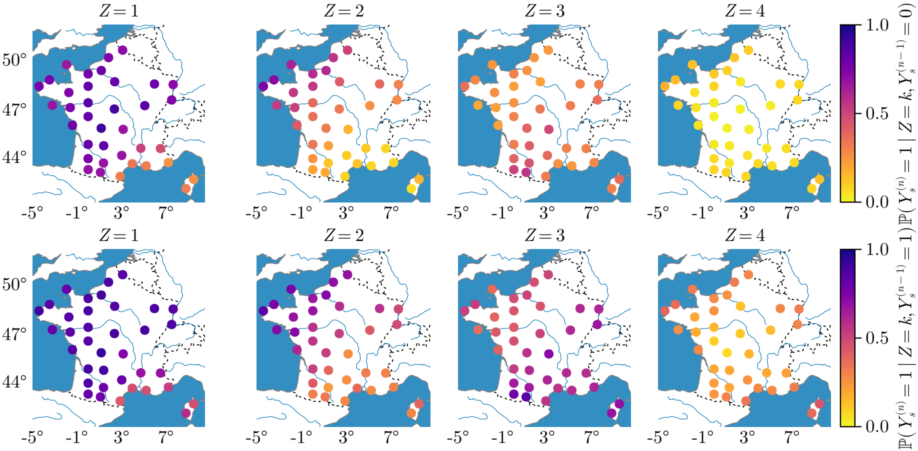
$$X_R(\cdot,n) \sim TMVN(0,\ C_R(\cdot),\ \boldsymbol{\ell}, \boldsymbol{u}),$$

$=$ $X_R(s,n)$ is marginally standard Gaussian.

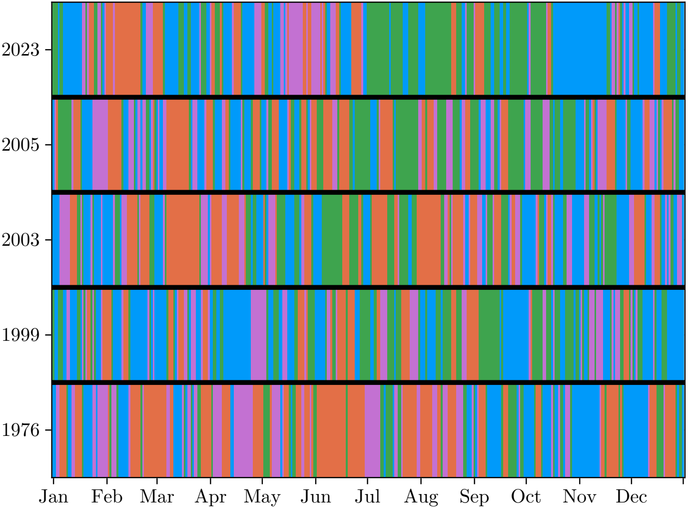
Back to rain intensity (like Obakrim et al, 2025) :
Given $X_R(s,n)$ and $Y_s^{(n)}$,

$$R_s^{(n)} = \begin{cases} \left(F^{(t_n)}_{z^{(n)},s}\right)^{-1}\left(\Phi_{T(s,n)}(X_R(s,n))\right), & Y_s^{(n)} = 1, \\ 0, & Y_s^{(n)} = 0. \end{cases}$$

Rain occurence - Parameters fitted from real data

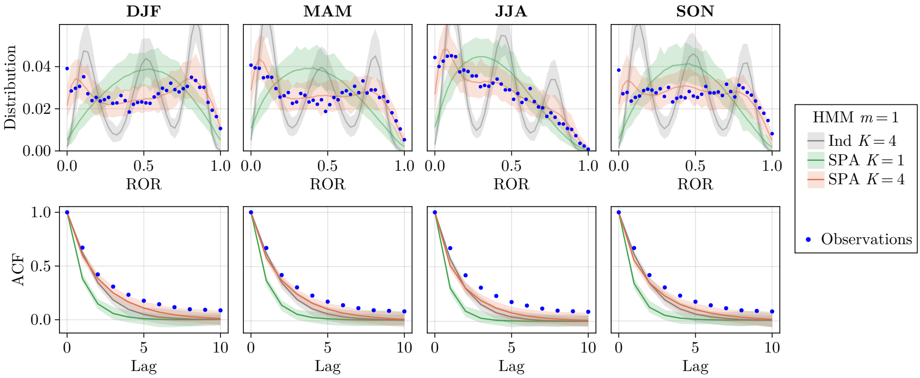


Rain occurence - Most likely sequence of states : interpretability



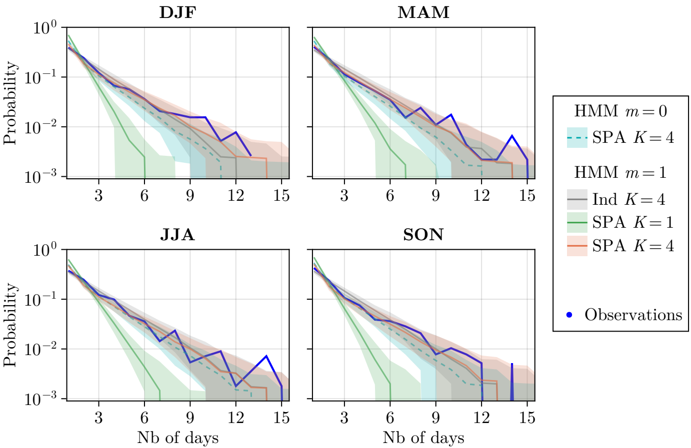
Rain occurence - Spatiotemporal evaluation

$$\text{Rain Occurrence Rate (ROR)} = \frac{\sum_{s\in\mathcal{S}} Y_s^{(t)}}{|\mathcal{S}|}$$



Rain occurence - Dry spells

Dry spell of Rain Occurrence Rate (ROR) $P_\ell = \mathbb{P}(ROR_t < 0.2, \ldots, ROR_{t+\ell} < 0.2)$



Entire generator with rain intensity

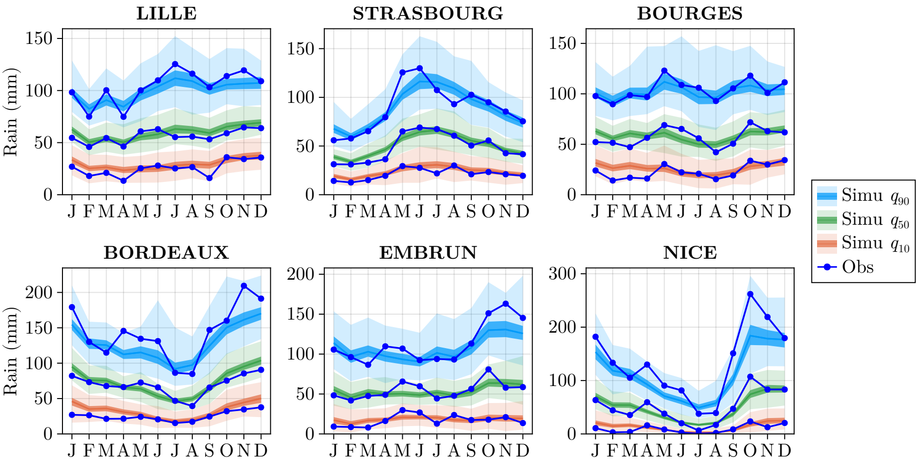


Figure: Monthly observed and simulated rainfall quantiles, for periodic EGPD parameters

- Cycle of seasonality well-represented

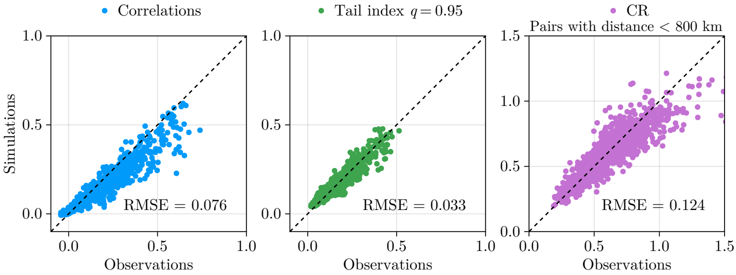
Entire generator with rain intensity

Make the generation of hidden state $Z^{(t)}$, then $Y_s^{(t)}|Z^{(t)}$, then some rain quantity $R_s^{(t)}|Z^{(t)}, Y_s^{(t)}$ and the final rain amount is NOT JUST $R \times Y$.

Spatial continuity ratio (Wilks, 1998)

For two sites s_k and s_ℓ, the **spatial continuity ratio** is

$$\mathrm{CR}(s_k, s_\ell) = \frac{\mathbb{E}\left[R_{s_\ell,t} > 0 \,\middle|\, R_{s_k,t} = 0\right]}{\mathbb{E}\left[R_{s_\ell,t} > 0 \,\middle|\, R_{s_k,t} > 0\right]}.$$



Conclusion

What has been done

- Seasonal Multisite rain occurrence model with hidden weather regimes
- No restriction on the distance between stations
- Evaluation of rain occurence with the spatiotemporal indicator ROR
- Added rain intensity "a posteriori" with marginals depending on the state, common covariance structure

Conclusion

What has been done

- Seasonal Multisite rain occurrence model with hidden weather regimes
- No restriction on the distance between stations
- Evaluation of rain occurence with the spatiotemporal indicator ROR
- Added rain intensity "a posteriori" with marginals depending on the state, common covariance structure

What is left to do

- Add other weather variables

Conclusion

What has been done

- Seasonal Multisite rain occurrence model with hidden weather regimes
- No restriction on the distance between stations
- Evaluation of rain occurence with the spatiotemporal indicator ROR
- Added rain intensity "a posteriori" with marginals depending on the state, common covariance structure

What is left to do

- Add other weather variables

Thank you ! Any questions ?