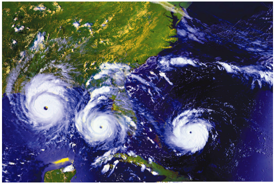
A frailty-contagion model for multi-site hourly precipitation driven by atmospheric covariates

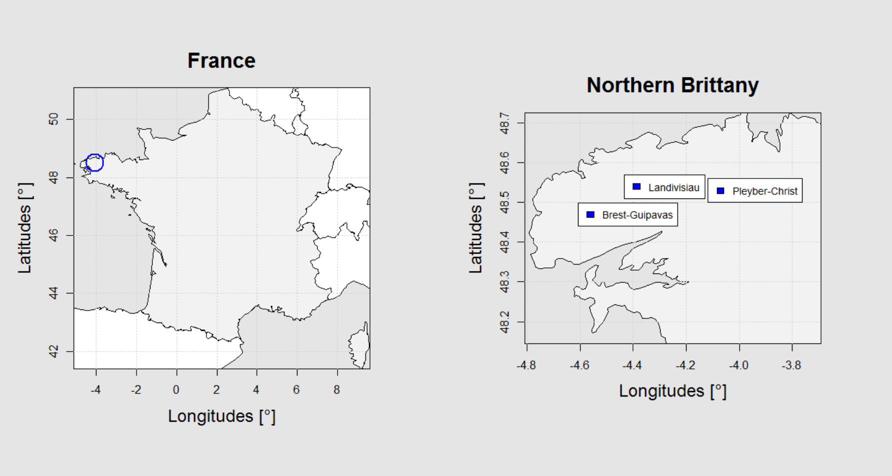
Erwan Koch (joint work with Philippe Naveau)

ETH Zurich, Department of Mathematics, RiskLab

Workshop on stochastic weather generators
Vannes
May 18, 2016



Motivation



Three weather stations in northern Brittany (France) with hourly winter precipitation records from 1995 to 2011.

Outline

1. Short review of precipitation generators
2. Heteroscedastic multi-site model
3. Inference method and simulation study
4. Application to hourly precipitation in Brittany
5. Conclusion and perspectives

Some seminal papers about single-site models

Parametric methods

Two **statistically independent** components:

- **The occurrence process**: Markov chains (Katz, 1977). In most cases a two-state, first-order Markov chain (Richardson, 1981).
- **The intensity process**: exponential distributions (Richardson, 1981), gamma distributions (Stern and Coe, 1984) or some exponential mixtures (Woolhiser and Pegram, 1979).

Some seminal papers about multi-site models

Parametric: purely statistical models

- Seminal paper by Wilks (1998) which extends Katz (1977). Models the **spatial correlation** of both precipitation occurrence and intensity using **latent multivariate normal distributions**.
- Extended in Wilks (2009) and Kleiber *et al.* (2012). Kleiber *et al.*'s (2012) approach departs from that of Wilks (2009) by embracing models for occurrence and intensity in a GLM framework.
- This category also includes generalized chain-dependent processes (Zheng and Katz, 2008; Zheng *et al.*, 2010) and copula-based approaches (Bardossy and Pegram, 2009).

Drawbacks

- Generally, independence between occurrence and intensity.
- They do not really take large scale conditions into account.
- Extreme values are overlooked in both sides of the precipitation distribution (zeros and very large values).

Some seminal papers about multi-site models

Parametric: Weather-state models

- Use available atmospheric data.
- Try to represent some physical processes originating precipitation.

Important case of hidden Markov models for occurrence (Hughes and Guttorp, 1994, 1999) and intensity (Ailliot *et al.*, 2009 ; Charles *et al.*, 1999). Precipitations depend on a latent process following a Markov chain whose **transition probabilities vary under atmospheric conditions**.

Drawbacks

- Generally, independence between occurrence and intensity.
- More parameters than in purely statistical models $\Rightarrow$ estimation is difficult.
- Extreme values are overlooked in both sides of the precipitation distribution (zeros and very large values).

Summary

Literature

- To the best of our knowledge, **no (or only a few) hourly** precipitation generator(s) so far.
- Generally, occurrence and intensity considered as **independent** (Katz, 1977; Kleiber et al., 2012; . . .).
- Extreme values are overlooked:
  - Difficulty to reproduce the **large precipitation amounts**.
  - Difficulty to reproduce the **persistence of dry periods**.

Link between occurrence and intensity

	Weather stations		
Intensity range (mm)	Brest-Guipavas	Landivisiau	Pleyber-Christ
0.2-0.8	0.63	0.65	0.68
0.8-1.4	0.74	0.75	0.79
1.4-2	0.81	0.82	0.82
2-4	0.90	0.91	0.91
> 4	0.94	0.94	0.94
Chi-Square test p-value	8.40×10^{-64}	4.57×10^{-59}	4.39×10^{-43}

Independence test

- Clear increasing relationship.
- The assumption of independence between occurrence and intensity is clearly unrealistic.

The model

Let $P_{m,t}$ $(m = 1, \ldots, M; t = 1, \ldots, T)$ be the precipitation amount at station m recorded during the t^{th} hour.

The model

$$P_{m,t} = \begin{cases} \mathbf{B}_m^{'} \mathbf{P}_{t-1} + \epsilon_{m,t} & \text{if } \mathbf{B}_m^{'} \mathbf{P}_{t-1} + \epsilon_{m,t} \geq u \\ 0 & \text{if } \mathbf{B}_m^{'} \mathbf{P}_{t-1} + \epsilon_{m,t} < u \end{cases},$$

where:

- $\mathbf{B}_m$ is a vector of autoregressive coefficients.
- The $\epsilon_{m,t}$ are independent and satisfy, conditionally to an observable common factor $\mathbf{F}_t$,

$$\epsilon_{m,t} \sim \mathcal{N}\left(0, \exp(\theta^{'} \mathbf{F}_t)\right).$$

- $\mathbf{F}_t$ is a vector of d atmospheric explanatory variables at time t.
- The threshold u is positive.

Maximum likelihood inference

We denote by B the matrix composed of the rows $\mathbf{B}_m^{'}$.

Maximum likelihood inference

The log-likelihood function denoted $L_u(B, \boldsymbol{\theta})$ for a given u can be written

$$L_u(B, \boldsymbol{\theta}) = \sum_{t=2}^{T} \sum_{m=1}^{M} \left\{ \mathbb{I}_{\{P_{m,t} \geq u\}} \ln \left[\frac{1}{\exp(\boldsymbol{\theta}' \mathbf{F}_t)} \phi \left(\frac{P_{m,t} - \mathbf{B}_m^{'} \mathbf{P}_{t-1}}{\exp(\boldsymbol{\theta}' \mathbf{F}_t)} \right) \right] + \mathbb{I}_{\{P_{m,t} = 0\}} \ln \left[\Phi \left(\frac{u - \mathbf{B}_m^{'} \mathbf{P}_{t-1}}{\exp(\boldsymbol{\theta}' \mathbf{F}_t)} \right) \right] \right\},$$

where $\Phi(.)$ and $\phi(.)$ represent the distribution function and the probability distribution function of the standardized normal distribution, respectively.

The maximum likelihood estimator is computed numerically.

Simulation study

Sample size	100			1000		
Parameter values	Bias	Stdev	Relative error	Bias	Stdev	Relative error
$B_{11} = 0.65$	-2.08	17.71	-3.19	0.00	0.05	0.00
$B_{12} = -0.08$	-8.74	22.97	114.37	-0.01	0.06	0.07
$B_{13} = 0.11$	-13.30	30.62	-122.71	0.00	0.06	0.03
$B_{21} = 0.47$	-3.44	17.94	-7.37	0.01	0.05	0.02
$B_{22} = 0.25$	-5.57	13.83	-22.36	0.00	0.04	-0.01
$B_{23} = 0.02$	-7.76	20.43	-435.58	0.00	0.06	-0.09
$B_{31} = 0.22$	-8.64	27.02	-39.95	0.00	0.05	-0.01
$B_{32} = 0.10$	-7.17	21.43	-75.40	0.00	0.05	0.02
$B_{33} = 0.36$	-8.18	24.75	-22.65	-0.01	0.05	-0.03
$\theta_0 = 30.63$	-12.36	146.13	-0.40	0.19	1.54	0.01
$\theta_1 = 0.07$	0.07	0.36	0.96	0.00	0.01	-0.01
$\theta_2 = 0.03$	-0.06	0.14	-0.27	-0.07	0.00	0.01
$\theta_3 = 0.03$	0.03	0.08	0.92	0.00	0.00	0.02

Inference assessment by simulations for two sample sizes of 100 and 1000. The empirical bias, standard deviation and relative error are derived from 100 independent replicas simulated from the model with parameters given in the first column.

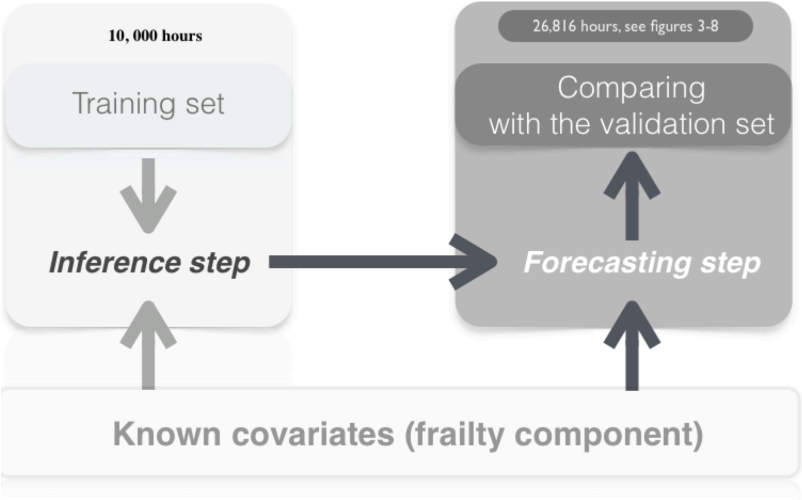
Estimates

Threshold: $u = 0.7$

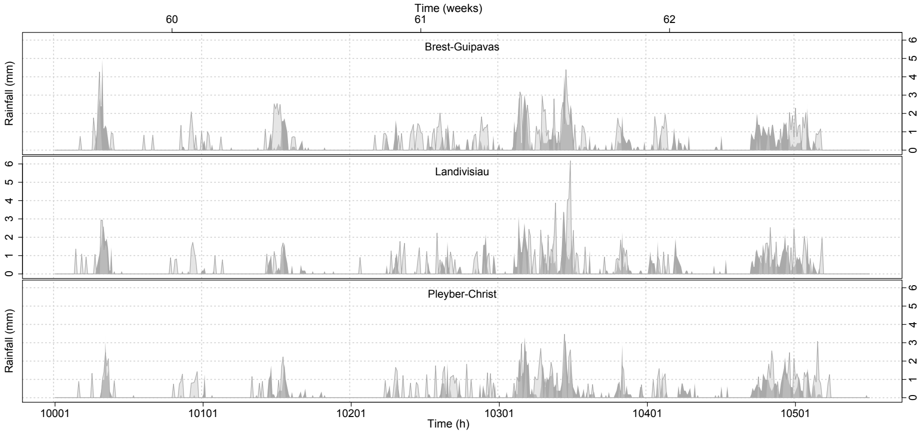
Matrix B	Brest-Guipavas	Landivisiau	Pleyber-Christ
Brest-Guipavas	0.65	-0.08	0.11
	[0.59 ; 0.74] (87)	[-0.15 ; 0.01] (86)	[0.02 ; 0.19] (89)
Landivisiau	0.47	0.25	0.02
	[0.41 ; 0.53] (81)	[0.17 ; 0.32] (92)	[-0.06 ; 0.09] (91)
Pleyber-Christ	0.22	0.10	0.36
	[0.16 ; 0.27] (77)	[0.02 ; 0.17] (82)	[0.30 ; 0.43] (80)

Atmospheric variables	Estimates	Confidence interval	Coverage probability
Regression intercept	$\hat{\theta}_0 = 30.626$	[28.02 ; 32.32]	80
Temperature	$\hat{\theta}_1 = 0.070$	[0.064; 0.076]	38
Seal level pressure	$\hat{\theta}_2 = -0.034$	[-0.037 ; -0.031]	94
Humidity	$\hat{\theta}_3 = 0.028$	[0.022; 0.036]	94

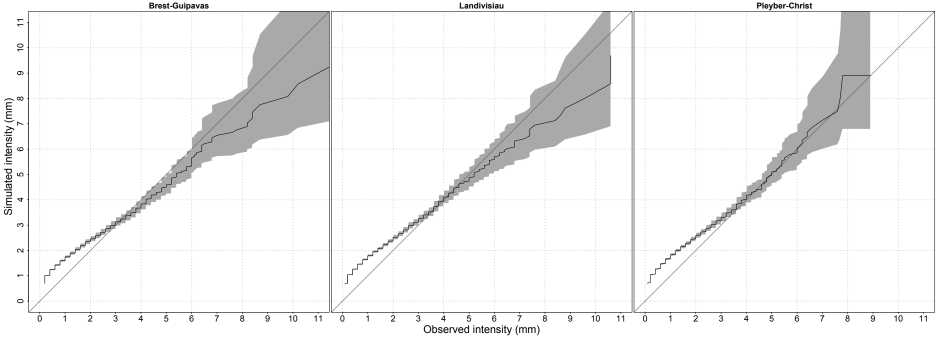
Forecast scheme



Comparison of the dynamics: out-sample time series

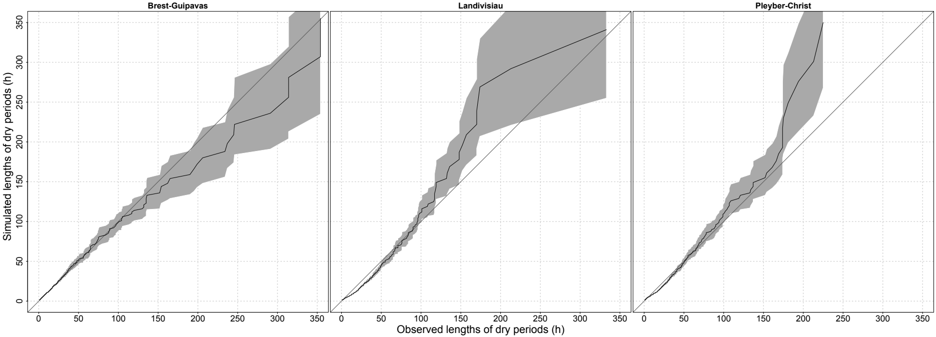


Out-sample intensities



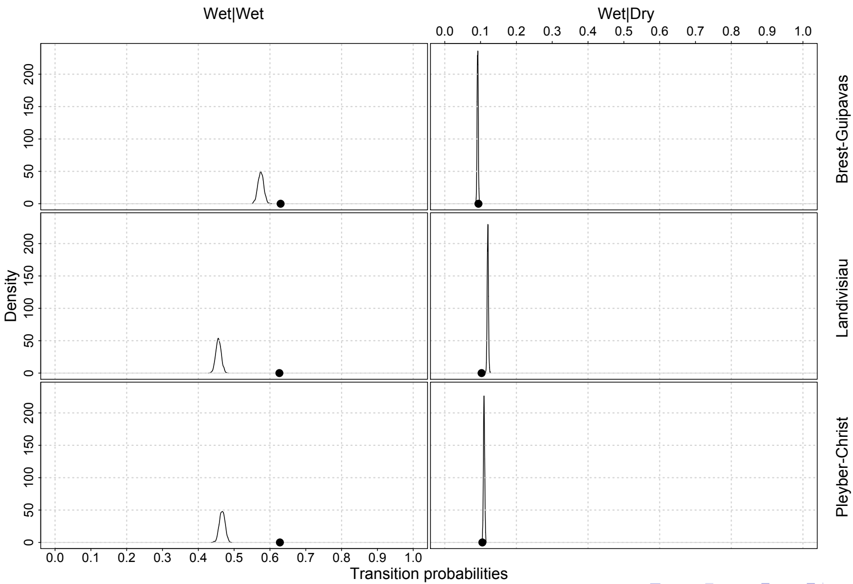
Quantile-quantile plot between observed rainfall amount (x-axis) and simulated one (y-axis) from the model. The gray color corresponds to the 98% confidence interval and the solid line to the median.

Out-sample dry periods

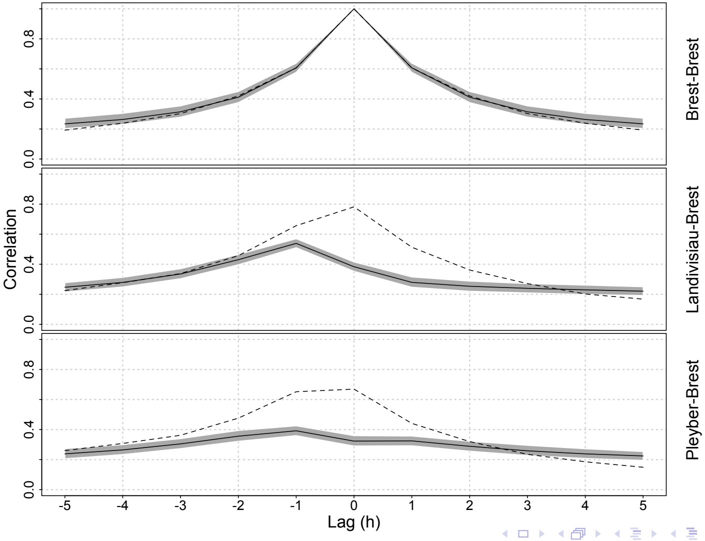


Quantile-quantile plot between observed dry periods length (x-axis) and simulated one (y-axis) from the model. The gray color corresponds to the 98% confidence interval and the solid line to the median.

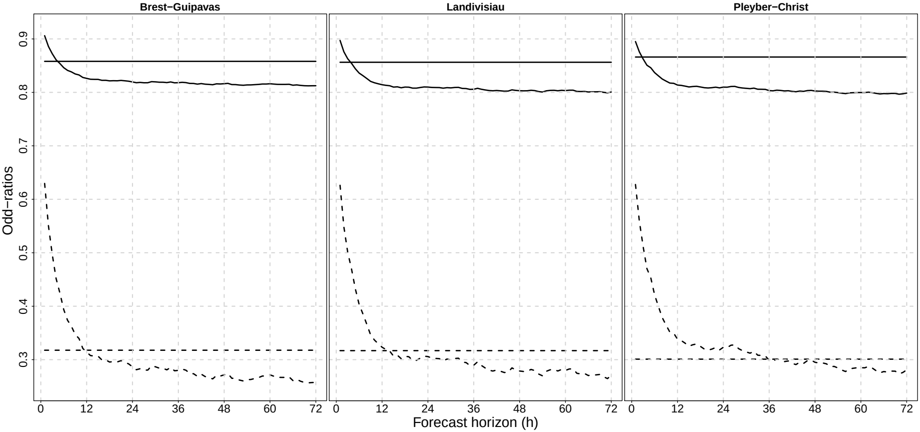
Out-sample transition probabilities



Out-sample correlations



Out-sample odd-ratios



Other experiences

Role of the frailty and contagion terms

- Model fitted **without the frailty component**: temporal dynamics, rainfall intensity and dry period persistence not reproduced accurately.
- Model fitted **without the contagion term**: performance quite good, although the persistence of wet periods seems to be underestimated.

Application of our model on daily precipitation

- Performance at least as good as in the hourly case.
- Comparison with the standard model by Wilks (1998):
  - Our model gives similar results concerning the general statistical properties (distributions of intensity and dry periods lengths, transition probabilities, cross-correlations, . . .).
  - Due to the absence of covariates, Wilks' model does not capture the dynamic structure, and especially the synchronicity between observed and simulated rainfalls.

Conclusion

- Simple model for **hourly** precipitation with 2 components:
  - **Auto-regressive part**: local scale effects.
  - **Heteroscedastic noise** based on covariates: large scale effects.
- Only **one equation, allowing to introduce dependence between occurrence and intensity.** Compared to models separating occurrence and intensity:
  - Easier to manipulate.
  - Easier to interpret.
  - More realistic.
- Reasonable results on French Brittany, **especially with regard to extremes**. Due to the fact that:
  - The model has the relevant structure.
  - The covariates have been well chosen.

Perspectives

- $\forall t$, $\{\varepsilon_{m,t}\}_{m=1,\ldots,M}$: components of a **general multivariate Gaussian vector** $\Longrightarrow$ second source of spatial dependence.
- $\forall t$, $\{\varepsilon_{m,t}\}_{m=1,\ldots,M}$: components of a **multivariate Student (or more generally elliptical) vector** $\Longrightarrow$ heavier tails (more realistic in the South of France for example).
- **Non-parametric relationship (like a Generalized Additive Model)** between the variance and the covariates.
- Stochastic model for the covariates.
- **Parametric structure on the coefficients** of matrix $B \Longrightarrow$ more tractable inference procedure in the case of M large.
- Analytical computation of $\mathbb{P}(P_{m,t} \leq u + x | P_{m,t-1} > u)$.

References

Allard, D. and Bourotte, M., 2015
Disaggregating daily precipitations into hourly values with a transformed censored latent Gaussian process.
Stochastic Environmental Research and Risk Assessment, 29 (2)

Kleiber, W., Katz, R., and Rajagopalan, B., 2012
Daily spatiotemporal precipitation simulation using latent and transformed gaussian processes.
Water Resources Research, 48 (1)

Koch, E. and Naveau, P., 2015
A frailty-contagion model for multi-site hourly precipitation driven by atmospheric covariates.
Advances in Water Resources, 78

Wilks, D. S., 1998
Multisite generalization of a daily stochastic precipitation generation model
Journal of Hydrology, 210 (1-4)

Wilks, D. S., 2009
A gridded multisite weather generator and synchronization to observed weather data
Water Resources Research, 45 (10)