



**HAL**  
open science

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

Erwan Koch, Philippe Naveau

► **To cite this version:**

Erwan Koch, Philippe Naveau. A frailty-contagion model for multi-site hourly precipitation driven by atmospheric covariates. *Advances in Water Resources*, 2015, 78, pp.145 - 154. 10.1016/j.advwatres.2015.01.001 . hal-01806250

**HAL Id: hal-01806250**

**<https://hal.science/hal-01806250>**

Submitted on 24 Jun 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

Erwan Koch\*

Philippe Naveau †

## Abstract

Accurate stochastic simulations of hourly precipitation are needed for impact studies at local spatial scales. Statistically, hourly precipitation data represent a difficult challenge. They are non-negative, skewed, heavy tailed, contain a lot of zeros (dry hours) and they have complex temporal structures (e.g., long persistence of dry episodes). Inspired by frailty-contagion approaches used in finance and insurance, we propose a multi-site precipitation simulator that, given appropriate regional atmospheric variables, can simultaneously handle dry events and heavy rainfall periods. One advantage of our model is its conceptual simplicity in its dynamical structure. In particular, the temporal variability is represented by a common factor based on a few classical atmospheric covariates like temperatures, pressures and others. Our inference approach is tested on simulated data and applied on measurements made in the northern part of French Brittany.

**Key words:** Common factor; Contagion; Precipitation simulators; Spatial-temporal dependence; Weather generators.

---

\*ISFA, CREST and ETH Zurich (Department of Mathematics, RiskLab). erwan.koch@math.ethz.ch

†LSCE (CNRS). philippe.naveau@lsce.ipsl.fr

# 1 Introduction

Stochastic weather generators (WG) are statistical models that aim at simulating quickly and realistically random sequences of atmospheric variables like temperatures, precipitation and wind speeds. Historically, weather generators started in hydrological sciences in the sixties and seventies. In 1962, Gabriel and Neumann proposed a Markov model for daily precipitation occurrences. Since then, a strong research effort in modeling precipitation distributions has sustained in the hydrological and statistical communities. This attention towards precipitation WGs can be explained by, at least, two different reasons. These stochastic simulations can be used in assessment studies, especially these linked to water resources managements. As one of the initial drivers in impact studies, simulated times series of precipitation can play a fundamental role in exploring some part of the sensitivity of floods, erosion and crops models. The second reason, very attractive to applied statisticians, is that modeling accurately precipitation distributions in space and time has always been an intriguing mathematical challenge. From a probabilistic point of view, a sequence of precipitation mixes two types of events (dry or wet) and the rainfall intensity represents a strong departure from the Gaussian territory (positive, skewed and sometime heavy tailed). Spatially and temporally, events can be strongly correlated within dry or wet episodes. Still today, the spatio-temporal dynamics of precipitation is difficult to model statistically. Numerous elaborate parametric approaches have been proposed in recent years (see, e.g. Furrer and Katz, 2008; Lennartsson et al., 2008; Vrac et al., 2007; Allard, 2012) (non-parametric approaches also exist but we will not discuss them here).

From a probabilistic point of view, most of daily precipitation generators decouple the occurrence and intensity processes. For example, Kleiber et al. (2012) first generated spatially and temporally correlated rainfall occurrences and then, at locations with positive precipitation, simulated their intensities. This strategy is classical in the WG literature (see, e.g. Katz, 1977; Richardson, 1981). It is mathematically convenient, in terms of inference, to frame the estimation scheme into a clear two step algorithm. But the hypothesis that the occurrence process can be simulated before the intensity process, i.e. independently of past, present or future rainfall amounts, could be challenged, especially at the hourly scale. If very large (small) rainfall amounts have been observed at time  $t$ , it seems more likely that a wet (dry) hour will follow at time  $t + 1$ . To quantify this naive reasoning, we need to introduce the example that will lead all our discussions in this article. We will work with winter (DJF) hourly precipitation recorded in the northern part of Brittany in France at three stations from 1995 to 2011, see Figure 1. In a nutshell, winter precipitation in this region are generally due to large scale perturbations coming from the Atlantic ocean.

Coming back to the question of simulating the occurrence process independently of the intensity, Table 1 shows that the probability of rainfall occurrence at time  $t + 1$  increases in function of the rainfall intensity at date  $t$  for each station. To couple the occurrence and intensity processes, a possible ap-

Intensity range (mm)	Weather stations in Fig. 1		
	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	$8.40 \times 10^{-64}$	$4.57 \times 10^{-59}$	$4.39 \times 10^{-43}$

Table 1: Rain occurrence probability at time  $t + 1$  given the rainfall intensity at  $t$ . Each row represents a specific range of hourly rainfall intensity. The Chi-Square p-values test the independence hypothesis.

proach, see e.g. Ailliot et al. (2009) and Kleiber et al. (2012) who discussed this issue on p.4, is to first generate a Gaussian random field, say  $Z_t$ . Secondly, a censoring mechanism applied to this field produces rainfall occurrences, i.e. wet events occur whenever  $Z_t$  is above a given threshold. Then, rainfall

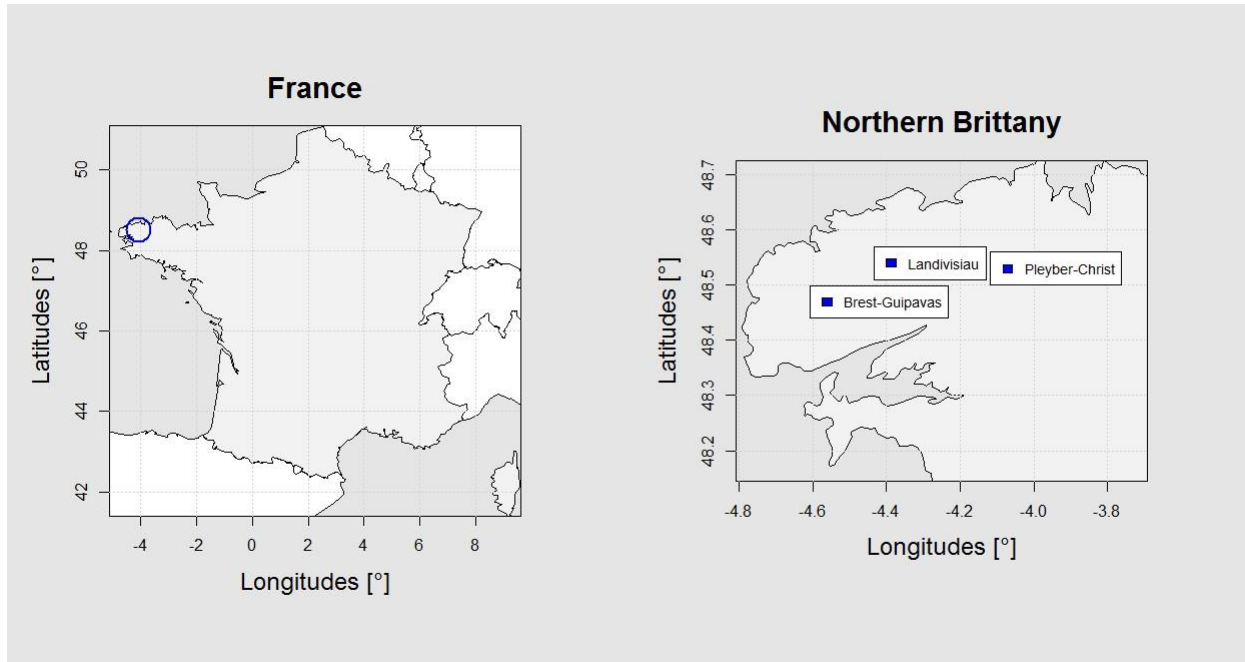


Figure 1: Three weather stations in northern Brittany (France) with hourly winter precipitation records from 1995 to 2011.

intensities are obtained by transforming the Gaussian values above the threshold into positive quantities (e.g., see Allard and Bourotte, 2013). The resulting rainfall intensity distribution depends on the choice of the transform function that may have different flavors (see, e.g. Wilks, 1998; Furrer and Katz, 2008; Lennartsson et al., 2008; Vrac et al., 2007; Allard, 2012). Concerning the correlation structure, the hidden process  $Z_i$ , by construction, drives both the intensity and occurrence processes that have become dependent. Another strategy (Serinaldi and Kilsby, 2014) is to directly fit a discrete-continuous at-site mixture distribution with two components (zero or rainfall amount). The parameters of these two components are then modelled by spatio-temporal processes (General Additive Model (GAM) or General Linear Model (GLM)).

This idea of working with an unobserved “seed” process is mathematically appealing but it adds some difficulties. In terms of statistical inference, it makes the estimation difficult. Concerning the interpretation, the latent process and its associated parameters are not necessary simple to explain. Although it has been a classical strategy to introduce hidden quantities in weather generators, for example see the abundant WG literature concerning hidden Markov structures (see reviews like Wilks and Wilby, 1999; Srikanthan et al., 2001; Ailliot et al., 2014), it would be a clear improvement, in terms of interpretation, if multi-site precipitation generators could bypass the hypothesis of hidden processes and directly model raw observations. This implies that additional information characterizing the dynamical structure of the weather system has to be integrated in a different way.

A well-maintained weather station recording hourly precipitation is rarely climatologically isolated in the sense that atmospheric variables like temperatures, pressures or others should be available in the same region. Most likely, the weather station itself may monitor this type of data. In this context, our plan is to develop the following two simple ideas in order to propose a new multi-site precipitation generator:

- (A) the main driver of current precipitation at a single location are precipitation recorded one hour early at all sites,
- (B) the second source of information comes from atmospheric variables like temperatures and others that jointly impact all sites by dynamically driving the variability of the residuals obtained from step (A).

Although resembling to existing concepts, this two step roadmap differs from the aforementioned studies by a few but important points. We will not represent the precipitation range by a finite mixture of distributions based on numerous weather types, each one representing different atmospheric conditions. Neither an hidden Gaussian process nor a function to transform a Gaussian variable into a Gamma type rainfall intensity are needed.

Our point (A) is not novel and it corresponds to the well-known idea of an auto-regressive process (see, for example the Markov chain in Katz, 1977). Here, we will use a basic multivariate auto-regressive representation (e.g., see Davis et al., 2014; Grimaldi et al., 2005). Point (B), modeling the residuals as a dynamical function of atmospheric variables, could be considered as more innovative in the WG literature. It is closely linked to the research developed in the statistical downscaling literature (e.g., see Maraun et al., 2010; Wilks, 2010, 2012). Downscaling approaches have mainly focused redon how to make the connection between circulation patterns and local atmospheric variables at the daily scale (e.g., see Vrac and Naveau, 2007; Flecher et al., 2010). Besides a different temporal scale (hourly in our case), our "spatial" component differs from downscaling. Our "common" signal, ordinary atmospheric variables averaged over our three sites in northern Brittany, corresponds to an information shared by all the sites, but it does not represent a regional feature over hundred of kilometers. The common point with downscaling is rather the idea that precipitation simulators performances can be improved if some appropriate explanatory variables can be injected in the statistical model.

The most interesting point of our article resides in combining ideas (A) and (B). From this coupling, a very simple model can handle precipitation in northern Brittany at the hourly scale, not only in terms of moderate rainfalls but also in terms of intense dry periods and even heavy precipitation. Before describing in detail our model, see Section 2, we need to say that our modeling strategy has been inspired by tools coming from two research fields not directly linked to geosciences: systemic risk in insurance and epidemic models.

Analyzing systemic risk in the financial system involves modeling the dependence between institutions. A first kind of dependence comes from common risk factors. These are called systematic factors, or frailties. For instance, the risk of life insurance contracts depends on the general uncertain increase of human lifetime, called longevity risk. As highlighted by the regulation since 2008, the second main source of dependence stems from contagion phenomenon between institutions. The contagion effect is due to the interconnections between banks and insurance companies by means of their debt and shares participations. Typically, the failure of a company will imply losses for its lenders, and maybe the failure of some of them. Then failure of some lenders of the lenders can occur, and so on. Originally, contagion models in finance come from the epidemic model introduced by Bailey (1953, 1957), Kendall (1956) and reintroduced in the so-called infectious model used in a static framework by Davis and Lo (2001). Specifications including both frailty and contagion effects are introduced among others in Frey and Backhaus (2003), Giesecke and Weber (2004, 2006), Azizpour et al. (2008), Gagliardini and Gouriéroux (2013).

Coming back to hourly precipitation data, our weather system in Brittany also contains both a common factor (frailty term) and a contagion term. The analogy with the financial system is the following. Each weather station corresponds to an institution. Our common factor represents large scale conditions that are, in some sense, exogenous to our system. For example, a storm front coming from the Atlantic ocean is not directly produced by the weather system observed in Brittany. It is similar to an exogenous shock affecting the whole financial system. Then once large scale processes are set then local physical processes are involved. These processes depend on the local characteristics. For instance, presence of mountains or flows can originate a privilegiate direction for thunder-storms propagation. It is some kind of contagion from one site to another. These local interactions between sites are the equivalent of cross-holdings at the origin of contagion in the financial system. Contrary to the usual assumptions used in finance, the common factor in climatology can be observable, covariates as temperatures, winds, pressures, and so on are also recorded at the weather stations and they represent valuable information that we have to take into account. Another link with financial statistics resides in the assumption that rainfall variability will be modelled by an heteroscedastic variance, a classical feature in econometrics. See e.g. the ARCH/GARCH model literature.

The remaining of the paper is organized as follows. Our heteroscedastic multi-site rainfall generator is detailed in Section 2. Then Section 3 presents our inference method based on maximum likelihood. Section 4 assesses our model's performance on simulations and summarizes our case study results dealing with the northern part of Brittany. Finally a discussion is given in Section 5.

## 2 A heteroscedastic multi-site rainfall generator

Denote  $M$  the number of weather stations and  $P_{m,t}$  the precipitation amount at station  $m$  recorded during the  $t^{\text{th}}$  hour. Our multi-site model based on (A) and (B) is defined as follows

$$P = \mathbf{B}' \mathbf{P}_{t-1} + \varepsilon_{m,t} \quad \text{if } \mathbf{B}' \mathbf{P}_{t-1} + \varepsilon_{m,t} \geq u, \quad (1)$$

where the scalar product

$$\mathbf{B}_m' \mathbf{P}_{t-1} = \beta_{m,1} P_{1,t-1} + \dots + \beta_{m,M} P_{M,t-1}$$

between the vector  $\mathbf{P}_{t-1} = (P_{1,t-1}, \dots, P_{M,t-1})'$  and  $\mathbf{B}_m = (\beta_{m,1}, \dots, \beta_{m,M})'$  represents the multivariate auto-regressive vector of order one that captures the dynamical local-scale effect, i.e. how the neighborhood behavior affects station  $m$  one hour later. The  $M \times M$  unknown auto-regressive coefficients  $\beta_{i,j}$  can be concatenated into a matrix, say  $\mathbf{B}$ , composed of the rows  $\mathbf{B}_m$ . For our Brittany example, large scale perturbations coming from the Atlantic ocean should make the matrix  $\mathbf{B}$  asymmetric. The Brest weather station should influence the two stations eastward, while the converse should not be true.

The random variable  $\varepsilon_{m,t}$  in (1) corresponds to the local variability driven by a few atmospheric variables and  $\varepsilon_{m,t}$  is simply modeled by a sequence of zero-mean Gaussian independent random variables. The non-stationarity comes from the standard deviation of  $\varepsilon_{m,t}$ ,  $\sigma_t > 0$ , that varies with time  $t$  in a log linear fashion

$$\ln \sigma_t = \theta' \mathbf{F}_t, \quad (2)$$

where  $\mathbf{F}_t$  represents a vector of  $d$  atmospheric explanatory variables at time  $t$ . The vector of scalars  $\theta$  corresponds to the unknown regression terms of the log-linear model in Equation (2). Concerning the spatio-temporal features of Equation (2), neither  $\theta$  and  $\mathbf{F}_t$  change from site to site, they only contain a spatially pooled information shared by all sites, and only  $\mathbf{F}_t$  varies with time. With a financial vocabulary, an economist to describe (1) and (2) will speak of a heteroscedastic model with a volatility driven by the "frailty term" (exogenous common factor)  $\mathbf{F}_t$ .

To ensure non-negative precipitation and produce dry events, Equation (1) contains the condition that rainfall only occur if the quantity  $\mathbf{B}_m' \mathbf{P}_{t-1} + \varepsilon_{m,t}$  is greater than some positive fixed threshold  $u$ . This could be loosely interpreted as saying that the rainfall generated at time  $t$ ,  $\mathbf{B}_m' \mathbf{P}_{t-1} + \varepsilon_{m,t}$ , has to be large enough to be recorded by the  $m^{\text{th}}$  weather station. This view emphasizes the fundamental role of having a temporally varying  $\sigma_t$ .

Imagine that all stations are dry at time  $t-1$ . All elements of the vector  $\mathbf{P}_{t-1}$  would then be equal to zero. In this case, the only way to produce large (small) rainfall at time  $t$  is to randomly generate a large (small)  $\varepsilon_{m,t}$ . This can be done by choosing a large (small) volatility,  $\sigma_t$ , or equivalently a large (small) frailty term  $\mathbf{F}_t$ . In other words, large scale conditions are captured by  $\mathbf{F}_t$  that drives the volatility of the system, and ultimately drives precipitation occurrences and strengths. This implies that heavy rainfall behavior, as well as dry period persistence, can be reproduced only if an adequate vector  $\mathbf{F}_t$  is chosen.

Compared to Kleiber et al. (2012), the dependence between occurrence and intensity is directly built in Equation (1). If it had rained a lot at time  $t-1$ , then  $\mathbf{P}_{t-1}$  is large, and it is very likely to have a wet hour at time  $t$  because of (1). In the opposite case, if  $\mathbf{P}_{t-1}$  is equal to zero, it is rather unlikely to have a wet hour at time  $t$ , unless the volatility is large. So, current rainfall occurrences are correlated with the intensity value observed one hour early.

Concerning the rain intensity distribution itself, our hypothesis that all  $\varepsilon_{m,t}$  are normally distributed does not imply that our simulated rainfall will follow a truncated Gaussian density, and consequently this does not mean that heavy tailed behavior cannot be simulated. The non-stationarity of  $\sigma_t$  explains

this phenomenon. At any time,  $\sigma_t$  can take a large value and therefore simulated precipitation resembles to a complex infinite mixture of truncated Gaussian random variables with a wide range of standard deviation. A consequence of this is that we don't need, at least in our example, to apply a power transform to go back and forth between the Gaussian world and the rainfall values like in Allard and Bourotte (2013) and Ailliot et al. (2009). Another reason is that  $\varepsilon_{m,t}$  does not represent raw precipitation but a type of increment between two consecutive hours, see Equation (1).

As emphasized in the previous paragraphs, the choice of  $\mathbf{F}_t$  is paramount in the overall capacity of our model to accurately simulate hourly precipitation. In order to limit overfitting, the dimension of  $d$  should not be too high and, to make our model useful, the type of covariates within  $\mathbf{F}_t$  should be easy to obtain in most rainfall applications. For example, the components of  $\mathbf{F}_t$  in our northern Brittany case are hourly temperatures, pressures at the sea level, and humidity, respectively. For each atmospheric variable, we spatially average hourly values recorded over our three sites to get  $\mathbf{F}_t$ . Before closing this section, we would like to emphasize that the choice of the threshold  $u$  plays an important role in accurately modeling the length of dry periods. The coming two sections will illustrate this point.

### 3 Inference

Given the vector  $\mathbf{F}_t$  and the threshold  $u$ , our inference scheme is based on maximizing the likelihood (ML) of model (1) with respect to the matrix of auto-regressive parameters  $\mathbf{B}$  and the regression coefficients  $(\theta_0, \dots, \theta_d)$  where  $\theta_0$  represents the intercept. The log-likelihood function denoted  $L_u(\mathbf{B}, \theta_0, \dots, \theta_d)$  for a given  $u$  can be written as (see the proof in the Appendix)

$$L_u(\mathbf{B}, \theta_0, \dots, \theta_d) = \sum_{t=2}^T \sum_{m=1}^M \mathbf{I}_{\{P_{m,t} \geq u\}} \ln \frac{1}{\sigma_t} \varphi \left( \frac{P_{m,t} - \mathbf{B}_m \mathbf{P}_{t-1}}{\sigma_t} \right) + \mathbf{I}_{\{P_{m,t} = 0\}} \ln \Phi \left( \frac{u - \mathbf{B}_m \mathbf{P}_{t-1}}{\sigma_t} \right), \quad (3)$$

where  $\Phi(\cdot)$  and  $\varphi(\cdot)$  represent the cumulative distribution function and the probability distribution function of the standardized normal distribution, respectively. The vector  $(\theta_0, \dots, \theta_d)$  appears throughout  $\sigma_t$ , see Equation (2). The indicator function  $\mathbf{I}_{\{P_{m,t} \geq u\}}$ , equal to one if  $P_{m,t} \geq u$  and zero otherwise, corresponds to our condition in (1) that generates either a dry or wet hour at station  $m$ . These indicator functions and the absence of a closed form for  $\Phi(\cdot)$  make the derivation of explicit ML estimates impossible, these values can only be obtained numerically. Our confidence intervals are derived by using a profile log-likelihood approach (Pawitan, 2001).

The above ML approach assumed that the threshold  $u$  has been correctly chosen. As our weather station precision is 0.2 mm, the scalar  $u$  is iteratively set to small values close to this instrument precision limit, say 0.2, 0.3, 0.4, 0.5, 0.6 and 0.7. To find the optimal  $u$  from this set, we arbitrarily chose a first guest, say .5 mm, and implement our ML inference that provides estimates for  $(\mathbf{B}, \theta_0, \dots, \theta_d)$ . For this set of ML estimates, we can simulate hourly precipitation for different values of  $u$  and then, choose the value of  $u$  minimizing the difference between the observed mean length of dry periods and the mean length of dry periods of the simulated precipitation. This process can be repeated with the new value  $u$  and we stop when the estimated threshold remains unchanged. To conclude on the choice of  $u$ , we note that, by construction, model (1) will never simulate values above zero but smaller than  $u$ . Hence, if our fixed  $u$  is little bit greater than the instrument precision of .2 mm, then very small but positive precipitation in Brittany are not considered in our ML optimization scheme, see Equation (3). This technical detail does not play an important role in our overall modeling strategy but allowing  $u$  to be greater than .2 mm improves substantially our capacity to reproduce dry period lengths.

### 4 Application

For our simulations and Brittany example, a learning set of 10, 000 hours and a validation set of 26, 816 hours are used to assess the performance of our estimation scheme. Figure 2 summarizes our inference

and validation scheme.

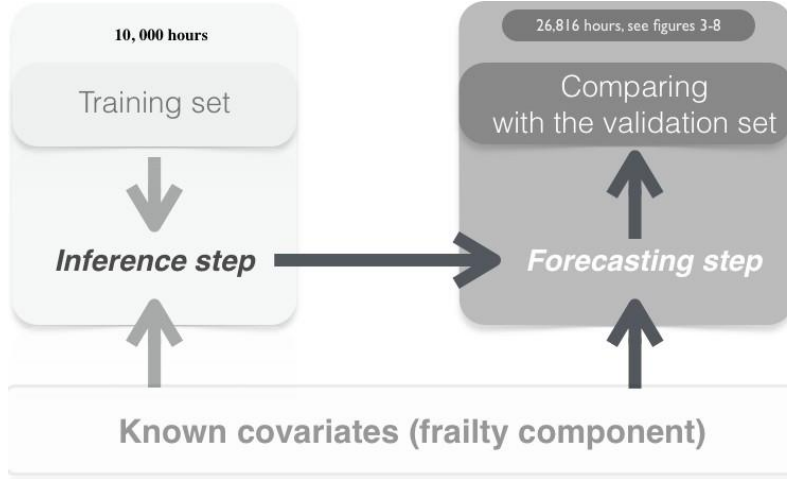


Figure 2: Inference and validation scheme: a learning set of 10, 000 hours is used to infer the parameters of our model. Then, assuming that the covariates are known in the future, we predict during the next 26, 816 hours and compare forecasted rainfall (so-called out-of-sample predictions) with precipitation recorded in Brittany. A few visual summaries of this forecast are shown in figures 3-8.

#### 4.1 Simulations

To test our inference scheme, we simulate 100 independent replicas for our model (1) with parameters that are chosen to be similar to our Brittany example, see the first column of Table 2. For two sample sizes (100 and 1000), the parameter estimates are derived by a classical ML approach based on Equation (3) and indicates that inferring 13 parameters with a small sample may be done with caution. The empirical mean, standard deviation and relative error derived from our 100 replicas are shown in Table 2. As expected, the sample length plays an important role in the inference. For a size of 100, the estimation can be difficult. In contrast, working with a sample size of 1000 provides accurate estimates. In practice, the sample size for hourly data is often long, i.e. more than 1000 hours. For our Brittany example, we have inferred our parameters with 10, 000 observations. In our simulations, the inference accuracy with a sample size of 10, 000 points is slightly better to the one with 1, 000 observations, except for  $\theta_0 = 30.63$  which is much improved (the bias and stdev become  $-0.005$  and  $0.59$ , respectively).

#### 4.2 Hourly precipitation in northern Brittany

The estimated auto-regressive coefficients  $\beta_{i,j}$  of the matrix  $\mathbf{B}$  with the 95% confidence intervals and coverage probability are displayed in Table 3. The first column makes it clear that the weather station of Brest has a strong influence in the two eastward stations. In particular, the station of Landivisiau has a larger coefficient with Brest (.47) than with itself (.25). In contrast, the two stations of Landivisiau and Pleyber-Christ, which are close to each other, do not have an impact on Brest. This pattern corresponds to the expected behavior for this region where rainfall comes from the Atlantic side, see Figure 1. This implies that Brest would be the first one to be impacted by a westerly front, and then the two eastward stations will be hit by the storm later on.

Concerning the three atmospheric variables in  $\mathbf{F}_t$  (temperatures, pressures at the sea level, and humidity) that drive our variability  $\sigma_t$ , Table 4 provides the respective estimated regression coefficients from (2) with their associated 95% confidence intervals and coverage probability. The values of  $\hat{\theta}_j$  for  $j = 1, 2, 3$  are small but their 95% confidence intervals do not contain zero. So, we keep these three explanatory variables in our rainfall generator.



Sample size	100			1000		
Parameter values	Bias	Stdev	Relative error	Bias	Stdev	Relative error
$\beta_{11} = 0.65$	-2.08	17.71	-3.19	0.00	0.05	0.00
$\beta_{12} = -0.08$	-8.74	22.97	114.37	-0.01	0.06	0.07
$\beta_{13} = 0.11$	-13.30	30.62	-122.71	0.00	0.06	0.03
$\beta_{21} = 0.47$	-3.44	17.94	-7.37	0.01	0.05	0.02
$\beta_{22} = 0.25$	-5.57	13.83	-22.36	0.00	0.04	-0.01
$\beta_{23} = 0.02$	-7.76	20.43	-435.58	0.00	0.06	-0.09
$\beta_{31} = 0.22$	-8.64	27.02	-39.95	0.00	0.05	-0.01
$\beta_{32} = 0.10$	-7.17	21.43	-75.40	0.00	0.05	0.02
$\beta_{33} = 0.36$	-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

Table 2: 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 (1) with parameters given in the first column.

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

Table 3: Estimated auto-regressive coefficients  $\beta_{ij}$  in (1) for the three weather stations plotted in Figure 1. The intervals represent the 95% confidence intervals and the number between brackets corresponds to the 95% coverage probability.

To visualize the predictive capacity of our model that has been fitted on the training set, we can generate synthetic hourly precipitation trajectories and compare them to the one kept in our validation period. For each weather station shown in Figure 1, Figure 3 compares observed rainfall (dark gray) with the simulated one (light gray) during 500 hours of the validation period, see Figure 2. Note that we have chosen a relative short period for the sake of visibility. However, the conclusions that can be drawn here are true whatever the period considered in the validation set (figures available upon request). In this exercise, we assume that the three atmospheric variables in  $\mathbf{F}_t$ , temperature, sea level pressure and humidity, are known at the regional scale.

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

Table 4: Estimated regression coefficients in (2) and corresponding 95% confidence intervals in function of our three explanatory atmospheric variables.

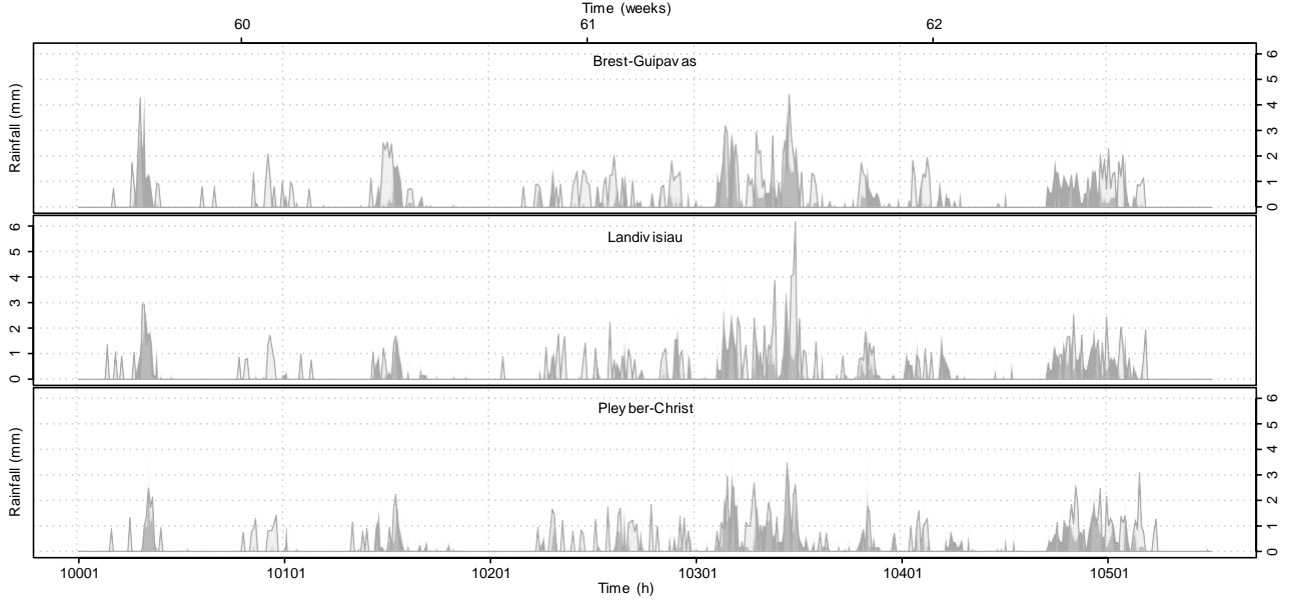


Figure 3: Each panel corresponds to one of our three weather stations shown in Figure 1. The dark and light gray color represent recorded precipitation and forecasted rainfall, respectively. Our predicted values are obtained on the validation period according to the scheme shown in Figure 2. The regional factor  $\sigma_t$  in (2) is the only quantity supposed to be known for the forecast. In particular, we don't use the precipitation recorded at time  $t$  to forecast the following hour or the following week.

The simulated rainfall appear to reproduce well the dynamical structure of real precipitation in the sense that clusters of simulated rainfall seem to be temporally and spatially synchronized with the observations. This indicates that combining the spatial factor  $\mathbf{F}_t$  with a multivariate auto-regressive structure drives accurately the dynamics of the system. Consequently, our model, conditionally on the common factor, cannot only be used as a rainfall generator but it can also carry out predictions. One can also notice in Figure 3 that real rainfall (left panel) have a jigsaw pattern. This is due to the recording device that is discrete in nature. Recorded precipitation amounts are added by increment of .2 mm (the precision of the instrument). This mechanical feature is not built in our statistical model defined by (1) because we view this as an undesirable characteristic of measured precipitation. This discrepancy explains the difference of "granularity" between the two panels. This phenomenon is particularly important for small and medium precipitation amounts. To illustrate this, we show the quantile-quantile plot between recorded and simulated rainfall intensities in Figure 4. The median and the 98% confidence interval in gray, obtained by parametric bootstrap (over 1000 simulated out-sample series), have been added. Rainfall amounts above 3 mm are well captured by our statistical model, this is particularly true for extreme intensities. Precipitation under 3 mm appear to be more difficult to reproduce and their intensities are slightly overestimated. This may be due to the nature of the recording process, the jigsaw pattern due to the instrument precision, and also to be the choice of our threshold  $u$  in model (1). The latter was optimized ( $u = 0.7$  mm) to capture accurately dry episodes. Figure 5 shows that this objective

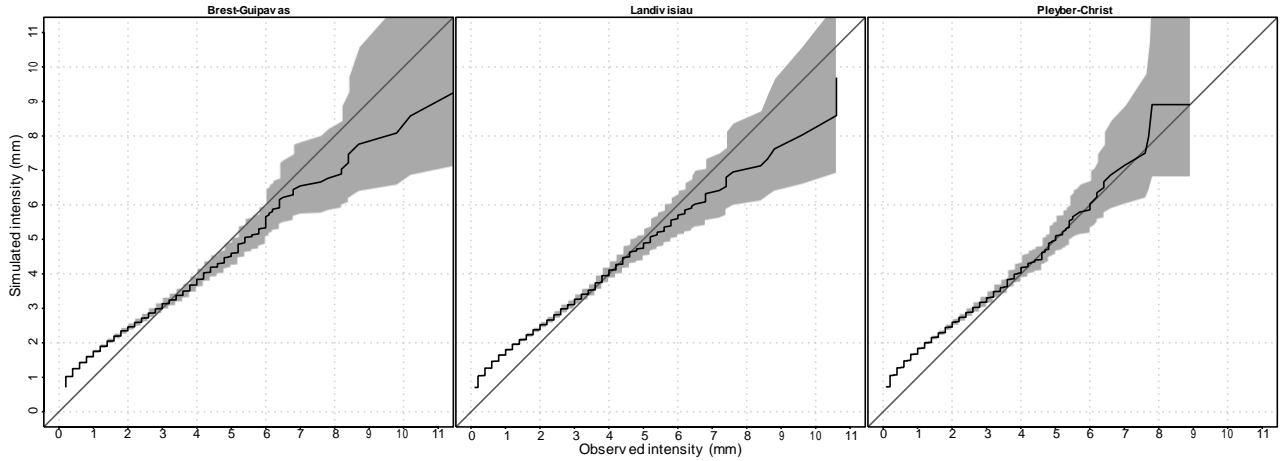


Figure 4: Out-sample quantile-quantile plot between observed rainfall amount (x-axis) and simulated one (y-axis) from model (1). Each panel corresponds to one of our three weather stations shown in Figure 1. The gray color corresponds to the 98% confidence interval and the solid line to the median.

has been basically reached. Both long and short dry episodes appear to be simulated accurately by our

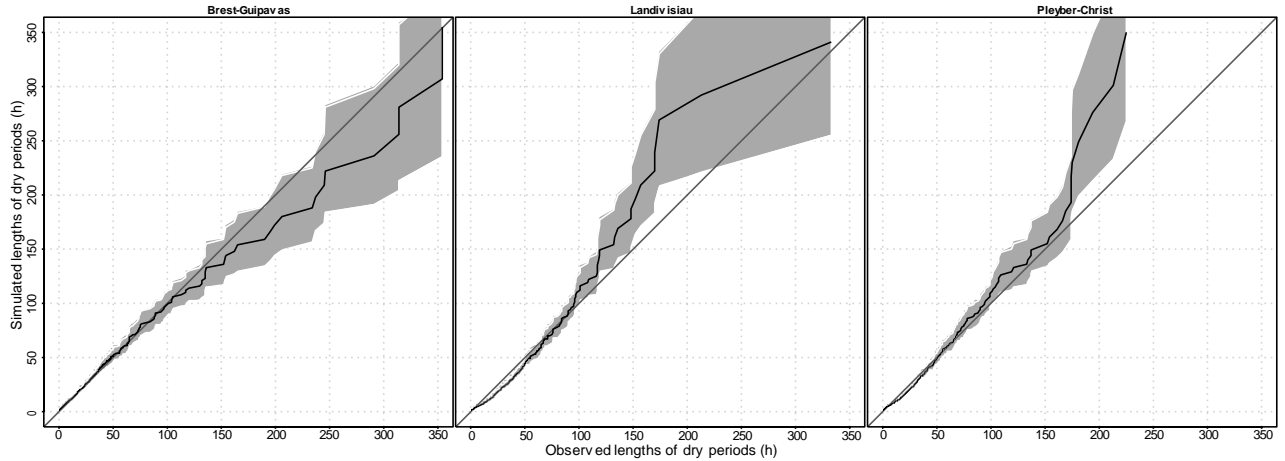


Figure 5: Out-sample quantile-quantile plot between observed dry periods length (x-axis) and simulated one (y-axis) from model (1). The gray color corresponds to the 98% confidence interval and the solid line to the median.

statistical model, especially at the Brest-Guipavas station.

To go one step further in analyzing our predictive out-sample, the probability of having a wet hour given that the preceding hour was also wet is displayed on the left panels of Figure 6, as well as the probability of moving from a dry hour to a wet one (right panels). Overall, these transition probabilities appear to give reasonable values, the black point corresponding to the observed estimate and the density to the distribution from simulated out-samples. This is particularly true for the dry to wet transition. Let us note that the other transition probabilities stem directly from these due to the relations  $P(\text{Wet}|\text{Wet}) + P(\text{Dry}|\text{Wet}) = 1$  and  $P(\text{Wet}|\text{Dry}) + P(\text{Dry}|\text{Dry}) = 1$ .

Concerning the temporal memory, the top panel of Figure 7 shows the out-sample auto-correlations at different time lags of one hour for the Brest-Guipavas station. This graph indicates that this local short term persistence (from one to five hours) is very well reproduced. The lower two panels focus on the out-sample cross-correlations between two pairs of stations, Landivisiau-Brest and Pleyber-Brest, respectively. These correlations are more difficult to reproduce: the observed ones (dotted lines) are

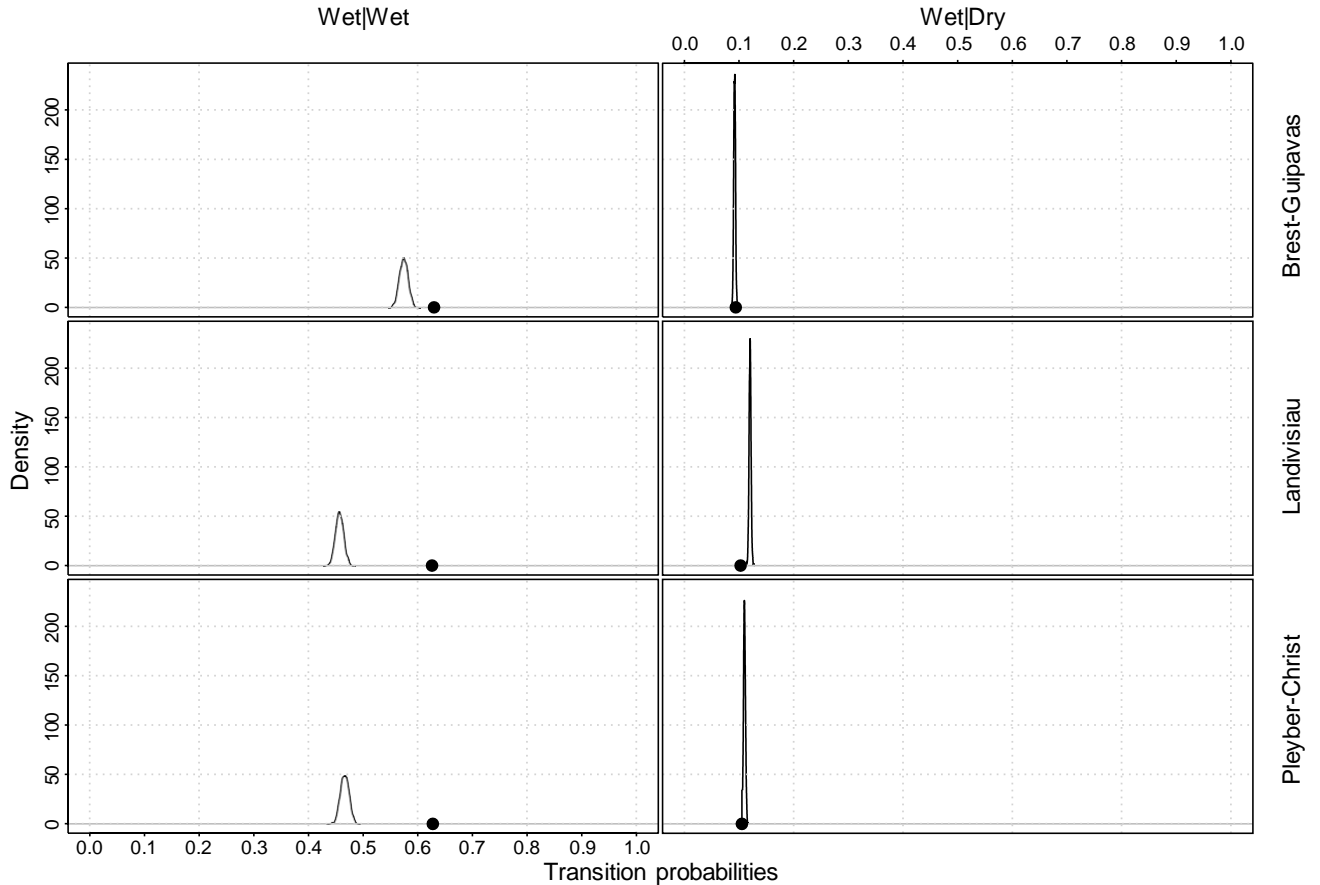


Figure 6: Each row represents a weather station. The left and right panels display the out-sample probability of having a wet hour given that the preceding hour was wet (left) and dry (right). The black points represent the observed estimates whereas the curves correspond to the density of estimates over 1000 out-sample simulated series.

above the ones computed on our out-sample. However, we capture the temporal asymmetry. The underestimation of correlations at lag  $-1$ ,  $0$  and  $1$  stems partly from the fact that the  $\{\varepsilon_{m,t}\}_{m=1,\dots,M}$  are i.i.d. This point will be further discussed in conclusion.

In order to analyze the prediction ability of our model, we have computed the following odd-ratios: the out-sample probability of predicting a dry hour while this hour is indeed dry and the out-sample probability of predicting a wet hour while this hour is indeed wet. In order to carry out a comparison, we have considered as a reference the simple model that gives as a prediction for a future horizon the value observed now. In Figure 8, we observe that from a certain value of the horizon, our model outperforms the naive prediction scheme. In some sense, this test allows to see approximately at which moment the frailty factor takes over the contagion term.

In our example, the frailty factor plays an important role in reproducing accurately temporal dynamics, rainfall intensity and dry period persistence. To test this (figures available upon request), our model was fitted without the frailty component and the aforementioned features were not adequately reproduced in such an instance. We have also fitted our model without the contagion term (figures available upon request). The performance is quite good, although the persistence of wet periods seems to be underestimated.

We would like to emphasize that figures 3-8 have been obtained on a validation set, totally different from the training one used to fit the model. This is reassuring and, albeit a few adjustments or extensions, we can expect that the basic ideas developed here could be relevant for other regions. These aspects are discussed in our conclusion.

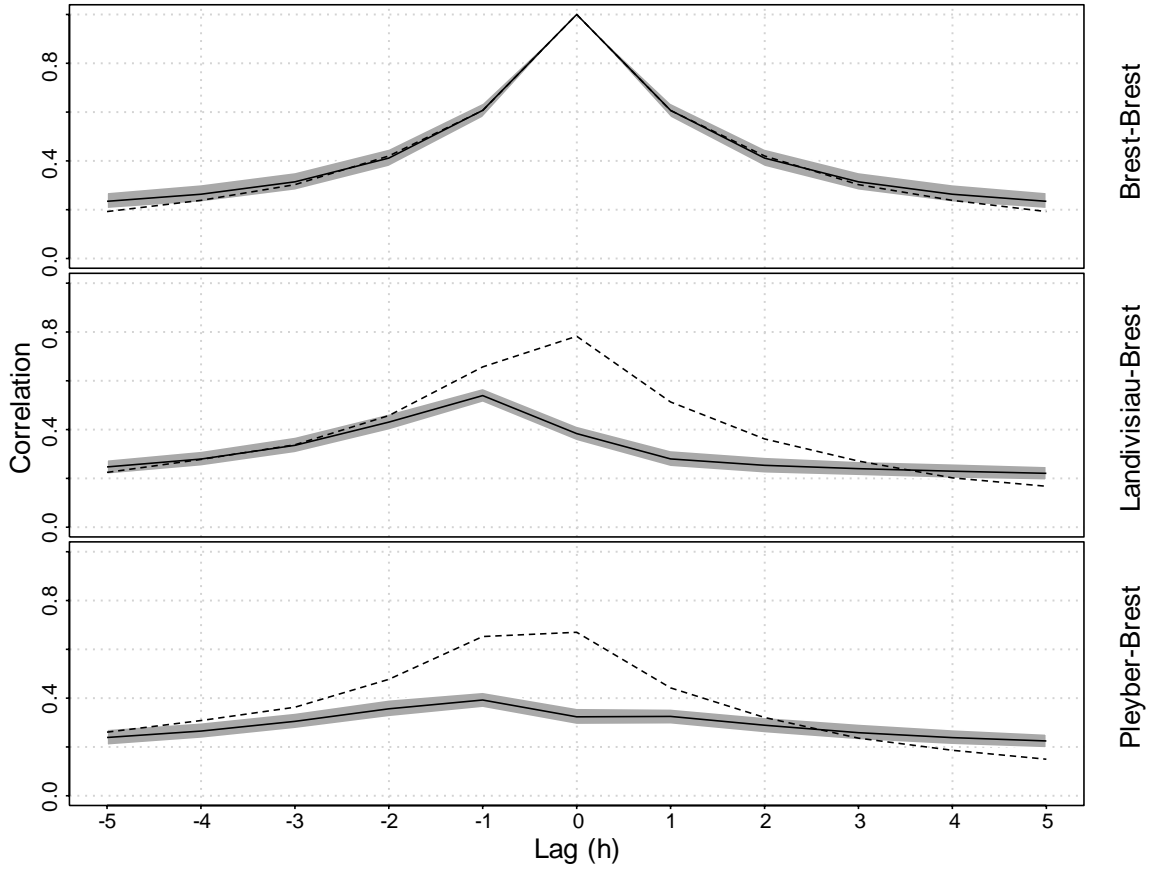


Figure 7: These panels correspond to out-sample correlations at different lags. On top, the auto-correlation at Brest-Guipavas. On the middle, the cross-correlation between Landivisiau and Brest-Guipavas and on bottom the cross-correlation between Pleyber-Christ and Brest-Guipavas. The gray color corresponds to the 98% confidence interval, the solid line to the median and the dashed one to the observed correlations.

Finally, we have also applied our model on daily precipitations measured at the same stations (see Figure 1). Our model's performance is at least as good as in the hourly case. Corresponding figures are available upon request. To give perspective, we compared our model 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. As already mentioned, these features are well reproduced by our model.

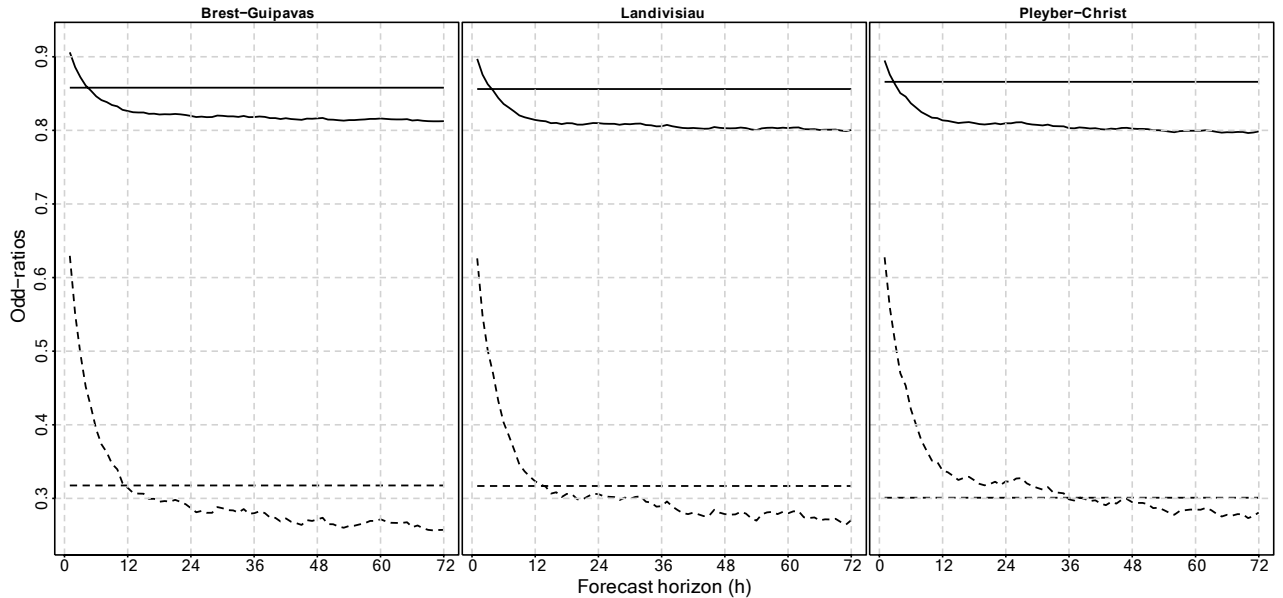


Figure 8: Each panel corresponds to one of our three weather stations shown in Figure 1. In solid are displayed the estimated probabilities to predict a dry hour while this hour is effectively dry. The constant line corresponds to the estimate obtained with our model. The curve corresponds to the model predicting for a given horizon the same value as now. In dashed, the same concerns the probabilities to predict a wet hour while this hour is effectively wet.

## 5 Concluding remarks

Our main objective was to show that a "simple" multivariate auto-regressive model with a heteroscedastic variance driven by a few well-chosen atmospheric covariates can provide an interesting blueprint to generate dry episodes, medium and heavy rainfall. From there, it is easy to extend this work by adding or modifying a few elements of our model. For example, we could imagine to replace the spatially independent random Gaussian noise  $\varepsilon_{m,t}$  by a multivariate Gaussian vector with a covariance matrix that could represent some spatial dependence among the different weather stations. One can even think about a multivariate student or elliptical distribution in order to provide heavier tails. This may be needed in regions with very heavy rainfall. Our northern Brittany example is known for frequent rainfall episodes but with rather low intensities, say compared to the South of France.

Another possible road for improvements would be to replace the linearity assumption in our regression model (2) by a non-parametric relationship (like a GAM, (e.g., Seraldi and Kilsby, 2014)) that will allow to capture some non-linear behavior. Although it was not really needed for our Brittany example, it is likely that the link between rainfall variability and temperatures could be more complex than our linear model for other regions.

It is also true that our multi-site statistical model is not a pure precipitation weather generator in the sense that we need more than precipitation data to fit and run our model. A few atmospheric covariates are necessary to drive our rainfall variability (for simulation of temperature time series, see e.g. Huong Hoang et al. (2009) and Hoang et al. (2011)). This limitation could be viewed as an advantage in the context of climate change studies. By letting our spatially averaged atmospheric covariates being driven by a numerical model, we could explore how precipitation change under different forcings. This is obviously related to downscaling themes and it would be interesting to pursue this path in future work. In this context, selecting appropriate covariates could be even more relevant, especially for large regions where scale indices like the North Atlantic Oscillation (NAO) Index or the El Niño Southern Oscillation (ENSO) Index could be useful.

One possible drawback of our approach may reside in our small number of sites. Modeling tens or

even hundreds of stations, instead of three, will lead to computational difficulties because of the  $M \times M$  size of the matrix  $\mathbf{B}$ . This could be solved by imposing a parametric structure on the  $\beta_{i,j}$ , for example they could decrease with the distance with respect to other stations or even set to zero for far apart stations. Finding a suitable distance is not trivial and should depend on orographic and other physical features.

To conclude, there are many different ways to fine tune and extend our approach. Basically, this will strongly depend on the application at hand and more research is needed to clearly see if this framework can be generalized.

## **Acknowledgements**

We thank Pierre Ribereau, Christian Y. Robert and the participants at the 12th International Meeting on Statistical Climatology and the ISI World Statistics Congress 2013 for their useful comments and suggestions. The authors have been financially supported by the different projects: ANR McSim, MIRACCLE-GICC, LEFE-Multirisk and Extremoscope. Erwan Koch would like to thank RiskLab at ETH Zurich and the Swiss Finance Institute for financial support.

## A Likelihood computation

*Proof.* We consider that the common factor's path  $\mathbf{F}_t$  ( $t = 1, \dots, T$ ) is given. Therefore we omit  $\mathbf{F}_t$  in the following. Let us denote by  $I_{t-1}$  the information available at time  $t - 1$ ,  $I_{t-1} = (P_{1,t-1}, \dots, P_{M,t-1}, P_{1,t-2}, \dots, P_{M,t-2}, \dots)$  and by  $h$  the density function. We can write by using a sequential argument that

$$h(P_{1,t}, \dots, P_{M,t}; t \in 1, \dots, T) = \prod_{t=2}^T h(P_{1,t}, \dots, P_{M,t} | P_{1,t-1}, \dots, P_{M,t-1}). \quad (4)$$

Since the  $\varepsilon_{m,t}$ ,  $m = 1, \dots, M$  are i.i.d. (conditionally on  $\mathbf{F}_t$ ), the variables

$$\mathbf{B}_m' \mathbf{P}_{t-1} + \varepsilon_{m,t}, m = 1, \dots, M$$

are independent. Thus it is the same for the variables

$$\mathbf{I}_{\mathbf{B}_m' \mathbf{P}_{t-1} + \varepsilon_{m,t} > u}, m = 1, \dots, M$$

and therefore for the product

$$P_{m,t} = \mathbf{B}_m' \mathbf{P}_{t-1} + \varepsilon_{m,t} \mathbf{I}_{\mathbf{B}_m' \mathbf{P}_{t-1} + \varepsilon_{m,t} > u}, m = 1, \dots, M.$$

Therefore the measurement equation given by model (1) gives that conditionally on  $(I_{t-1})$ , the variables  $P_{1,t}, \dots, P_{M,t}$  are independent, yielding

$$h(P_{1,t}, \dots, P_{M,t} | P_{1,t-1}, \dots, P_{M,t-1}) = \prod_{m=1}^M h(P_{m,t} | P_{1,t-1}, \dots, P_{M,t-1}). \quad (5)$$

Combining Equations (4) and (5), we obtain

$$h(P_{1,t}, \dots, P_{M,t}; t \in 1, \dots, T) = \prod_{t=2}^T \prod_{m=1}^M h(P_{m,t} | P_{1,t-1}, \dots, P_{M,t-1}). \quad (6)$$

Let us recall that

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

Due to this threshold mechanism,  $P_{m,t}$  is a mixture of a discrete random variable and a continuous one. Let us denote by  $h_d$  the density of the discrete part and by  $h_c$  the density of the continuous one. We have

$$h_c(P_{m,t} | P_{1,t-1}, \dots, P_{M,t-1}) = \frac{1}{\sigma_t} \varphi \left( \frac{P_{m,t} - \mathbf{B}_m' \mathbf{P}_{t-1}}{\sigma_t} \right)$$

and

$$\begin{aligned} h_d(P_{m,t} | P_{1,t-1}, \dots, P_{M,t-1}) &= \mathbb{P}[P_{m,t} = 0 | P_{1,t-1}, \dots, P_{M,t-1}] \\ &= \mathbb{P}[\mathbf{B}_m' \mathbf{P}_{t-1} + \varepsilon_{m,t} < u] \\ &= \mathbb{P}[\varepsilon_{m,t} < u - \mathbf{B}_m' \mathbf{P}_{t-1}] \\ &= \Phi \left( \frac{u - \mathbf{B}_m' \mathbf{P}_{t-1}}{\sigma_t} \right). \end{aligned}$$



Using Equation (6), the log-density is written

$$\begin{aligned}
& \ln [h(P_{1,t}, \dots, P_{M,t} | P_{1,t-1}, \dots, P_{M,t-1})] \\
&= \sum_{t=2}^T \sum_{m=1}^M \ln [hc(P_{m,t} | P_{1,t-1}, \dots, P_{M,t-1}) \mathbf{I}_{\{P_{m,t} \geq u\}} + hd(P_{m,t} | P_{1,t-1}, \dots, P_{M,t-1}) \mathbf{I}_{\{P_{m,t} = 0\}}] \\
&= \sum_{t=2}^T \sum_{m=1}^M \mathbf{I}_{\{P_{m,t} \geq u\}} \ln [hc(P_{m,t} | P_{1,t-1}, \dots, P_{M,t-1})] + \mathbf{I}_{\{P_{m,t} = 0\}} \ln [hd(P_{m,t} | P_{1,t-1}, \dots, P_{M,t-1})].
\end{aligned}$$

That finally yields the log-likelihood function

$$\begin{aligned}
& L_u(\mathbf{B}, \theta_0, \dots, \theta_d) \\
&= \sum_{t=2}^T \sum_{m=1}^M \mathbf{I}_{\{P_{m,t} \geq u\}} \ln \left[ \frac{1}{\sigma_t} \varphi \left( \frac{P_{m,t} - \mathbf{B}_m' \mathbf{P}_{t-1}}{\sigma_t} \right) \right] + \mathbf{I}_{\{P_{m,t} = 0\}} \ln \left[ \Phi \left( \frac{u - \mathbf{B}_m' \mathbf{P}_{t-1}}{\sigma_t} \right) \right],
\end{aligned}$$

completing the proof. □

## References

- Ailliot, P., Allard, D., Monbet, V., and Naveau, P. (2014). Stochastic weather generators: an overview of weather type models. *Submitted*.
- Ailliot, P., Thompson, C., and Thomson, P. (2009). Space time modeling of precipitation using a hidden Markov model and censored Gaussian distributions. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 58(3):405–426.
- Allard, D. (2012). Modeling spatial and spatio-temporal non Gaussian processes. In Porcu, E., Montero, J.-M., and Schlather, M., editors, *Advances and Challenges in Space-time Modelling of Natural Events*, Lecture Notes in Statistics, pages 141–164. Springer Berlin Heidelberg.
- Allard, D. and Bourotte, M. (2013). Disaggregating daily precipitations into hourly values with a transformed censored latent Gaussian process. *Stochastic Environmental Research and Risk Assessment*, Submitted.
- Azizpour, S., Giesecke, K., et al. (2008). Self-exciting corporate defaults: contagion vs. frailty. *Manuscript, Stanford University*.
- Bailey, N. T. (1953). The total size of a general stochastic epidemic. *Biometrika*, 40(2):177–185.
- Bailey, N. T. (1957). *The mathematical theory of epidemics*. Griffin London.
- Davis, M. and Lo, V. (2001). Modelling default correlation in bond portfolios. *Mastering risk*, 2:141–151.
- Davis, R. A., Zang, P., and Zheng, T. (2014). Sparse vector autoregressive modeling. *Submitted*.
- Flecher, C., Naveau, P., Allard, D., and Brisson, N. (2010). A stochastic daily weather generator for skewed data. *Water Resources Research*, 46(7).
- Frey, R. and Backhaus, J. (2003). Interacting defaults and counterparty risk: a Markovian approach. *University of Leipzig, Leipzig, Germany*.
- Furrer, E. and Katz, R. (2008). Improving the simulation of extreme precipitation events by stochastic weather generators. *Water Resources Research*, 44(12).
- Gabriel, K. and Neumann, J. (1962). A Markov chain model for rainfall occurrence at Tel Aviv. *Quarterly Journal of the Royal Meteorological Society*, 88:90–95.
- Gagliardini, P. and Gouriéroux, C. (2013). Correlated risks vs contagion in stochastic transition models. *Journal of Economic Dynamics and Control*, 37(11):2241–2269.
- Giesecke, K. and Weber, S. (2004). Cyclical correlations, credit contagion, and portfolio losses. *Journal of Banking & Finance*, 28(12):3009–3036.
- Giesecke, K. and Weber, S. (2006). Credit contagion and aggregate losses. *Journal of Economic Dynamics and Control*, 30(5):741–767.
- Grimaldi, S., Serinaldi, F., and Tallorini, C. (2005). Multivariate linear parametric models applied to daily rainfall time series. *Advances in Geosciences*, 2(2):87–92.
- Hoang, T., Dacunha-Castelle, D., and Benmenzer, G. (2011). Estimation of a diffusion model with trends taking into account the extremes. Application to temperature in France. *Environmetrics*, 22(3):464–479.

- Huong Hoang, T., Parey, S., and Dacunha-Castelle, D. (2009). Multidimensional trends: The example of temperature. *The European Physical Journal-Special Topics*, 174(1):113–124.
- Katz, R. (1977). Precipitation as a chain-dependant process. *Journal of Applied Meteorology*, 16:671–676.
- Kendall, D. G. (1956). Deterministic and stochastic epidemics in closed populations. In *Proceedings of the third Berkeley Symposium on Mathematical Statistics and Probability*, volume 4, pages 149–165.
- Kleiber, W., Katz, R., and Rajagopalan, B. (2012). Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes. *Water Resources Research*, 48(1).
- Lennartsson, J., Baxevani, A., and Chen, D. (2008). Modelling precipitation in Sweden using multiple step Markov chains and a composite model. *Journal of Hydrology*, 363(1):42–59.
- Maraun, D., Wetterhall, F., Ireson, A., Chandler, R., Kendon, E., Widmann, M., Brienen, S., Rust, H., Sauter, T., Themeßl, M., et al. (2010). Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, 48(3).
- Pawitan, Y. (2001). *All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press.
- Richardson, C. (1981). Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research*, 17(1):182–190.
- Serinaldi, F. and Kilsby, C. G. (2014). Simulating daily rainfall fields over large areas for collective risk estimation. *Journal of Hydrology*, 512:285–302.
- Srikanthan, R., McMahon, T., et al. (2001). Stochastic generation of annual, monthly and daily climate data: A review. *Hydrology and Earth System Sciences Discussions*, 5(4):653–670.
- Vrac, M. and Naveau, P. (2007). Stochastic downscaling of precipitation: from dry events to heavy rainfalls. *Water Resources Research*, 43(7).
- Vrac, M., Stein, M., and Hayhoe, K. (2007). Statistical downscaling of precipitation through non homogeneous stochastic weather typing. *Climate Research*, 34:169–184.
- Wilks, D. (1998). Multisite generalization of a daily stochastic precipitation generation model. *Journal of Hydrology*, 210(1):178–191.
- Wilks, D. (2010). Use of stochastic weather generators for precipitation downscaling. *Wiley Interdisciplinary Reviews: Climate Change*, 1(6):898–907.
- Wilks, D. (2012). Stochastic weather generators for climate-change downscaling, part ii: multivariable and spatially coherent multisite downscaling. *Wiley Interdisciplinary Reviews: Climate Change*, 3(3):267–278.
- Wilks, D. and Wilby, R. (1999). The weather generation game : a review of stochastic weather models. *Progress in Physical Geography*, 23(3):329–357.