Interpretable Seasonal Hidden Markov Model for spatio-temporal stochastic rain generation in France

Emmanuel Gobet[1], David Métivier[1,3], and Sylvie Parey[2]

[1]CMAP, CNRS, École Polytechnique, Institut Polytechnique de Paris, Route de Saclay, Palaiseau, France
[2]EDF R&D, 6 quai Watier, 78401 Chatou Cedex, France
[3]MISTEA, Université de Montpellier, INRAE, Institut Agro, Montpellier, France

June 24, 2024

Abstract

We present a Stochastic Weather Generator described based on a multisite Hidden Markov Model (HMM) and trained with French weather stations data. It generates correlated precipitation, with a special focus on seasonality and the correct reproduction of the distribution of dry and wet spells. The hidden states are viewed as global weather regimes, e.g., dry all over France, rainy in the north, etc. The resulting model is fully interpretable; it can even approximately recover large-scale structures such as North Atlantic Oscillations. The model achieves very good performances, specifically in terms of extremes. Its architecture allows easy integration of other weather variables. We show an application where the model is trained on future climate scenarios, allowing easy comparison and interpretation with the historical data in terms of parameters evolution and extremes.

KEYWORDS: hidden Markov models; statistical estimation; stochastic generators; rainfall modelling

MSC2020: 62M05; 37H10; 62P12

1 Introduction

Context. The current climate change context necessitates for industrials a careful analysis of the resilience of their assets under future climate conditions, to anticipate possible adaptation needs. Such

Acknowledgement: The authors thank Valérie Monbet and Mathieu Vrac for the feedbacks on an early version of this work, as well as Marc Lavielle for useful discussions.
This work benefited from the support of the *Chaire Énergies Durables* (CEA-EDF-École polytechnique) and of the *Chaire Stress Test: Risk management and Financial Steering* (BNP Paribas, École polytechnique) from the Fondation de l'École polytechnique.

analyses imply the estimation of future extreme hydrometeorological conditions, for example, the frequency of long-lasting dry spells for hydropower or nuclear generation. Such needs have been highlighted by some parliamentary missions in France [Christophe and Pompili, 2018], asking to quantify the impact of hydro-stress on nuclear power generation. Quantifying the impacts of future hydrometeorological conditions on crops is also important for farmers' strategies (see [Pascual et al., 2017] and references therein).

Climate models, global as well as regional, are powerful tools to simulate the climate system and project its possible evolution under different forcing scenarios. However, because they are computationally expensive, they do not allow performing numerous simulations. In practice, these weather scenarios, once generated, are made available on a public repository, see for instance the French national project DRIAS [Soubeyroux et al., 2021] (where one can find around thirty projections, see Section 8 for details); however, for more accurate risk assessment purposes, more scenarios may be needed and one may face a problem of data augmentation (how to resample easily new scenarios). Furthermore, even though progress has been made in terms of spatial resolution and process modeling, they still do not accurately enough reproduce local extremes. Therefore, stochastic weather generators are still widely used in impact studies. The recent IPCC working group 1 report of the 6th assessment [Arias et al., 2021] clearly states that: "Methodologies such as statistical downscaling, bias adjustment and weather generators are beneficial as an interface between climate model projections and impact modeling and for deriving user-relevant indicators". Whereas climate models represent the physics governing the climate system evolution, stochastic weather generators are calibrated to best reproduce the statistical properties of the climate variables, in terms of distribution, spatial and temporal correlations, or intervariable dependence for multivariate models.

Generating statistically coherent weather series in time and space is a hard problem. Mathematically speaking, they are multivariate time series that are far from being independent and identically distributed. In crude terms, today's weather is strongly influenced by what happened yesterday and also correlated with its surroundings. Moreover, the weather conditions change throughout the year and with climate change. Additionally, extreme events at both ends of the spectrum like extreme precipitation or intense heatwaves have to be well reproduced. The purpose of this paper is to build a parametric model able to quickly simulate a very large ensemble of multisite and mutually consistent weather variables so that very rare combinations can be reached.

Besides, weather simulation models are essential tools for climate stress testing, see [Ranger et al., 2022] for instance. By providing accurate forecasts and detailed scenarios, they enable decision-makers to anticipate climate challenges and take proactive measures to protect societies and ecosystems. In a world increasingly affected by climate change, the effective use of these models is essential to build resilience and ensure a sustainable future; thanks to the good interpretability properties of our model's parameters (unlike a pure generative model based on neural networks like in [Goodfellow et al., 2014]), we are able to parameterize precipitation scenarios by taking climate change factors into account, see subsection 8.3. Also see [Miloshevich et al., 2024] for a comparison between Stochastic Weather Generator and Deep Learning models.

Background literature. Many weather generators are devoted to the generation of precipitation time series, since precipitations are a crucial variable for many impact studies in agriculture or hy-

drology, for example, see [Wilks and Wilby, 1999, Ailliot et al., 2015a] for reviews. Stochastic Weather Generator's (SWG) development dates from the 1980s, with the model proposed by [Richardson, 1981] to generate long samples of precipitation, minimum and maximum temperature and solar radiation. Today, many approaches have been proposed.

The authors of [Cowpertwait et al., 2007] base their model on the generation of storm cells whose occurrence follows a Poisson process, during which rain cells occur as a secondary Poisson process. Other generators are based on meteorologically defined weather types (dry, wet or atmospheric variables). These types are identified through classification of the rainy and non-rainy days separately for each season, and the number of weather types is chosen according to a model selection criterion (BIC for instance). The authors of [Vrac et al., 2007] proposed such a model used for precipitation downscaling with weather types identified a priori through classification either of precipitation data or of exogenous atmospheric variables. Such inference work is delicate since not only one has to select the relevant weather types, but also one has to infer their stochastic properties (in addition to the stochastic properties of precipitation conditionally to weather types).

To circumvent these difficulties, Hidden Markov Models (HMM) introduce the weather types as latent variables [Ailliot et al., 2009, Sansom and Thomson, 2010]: the states form a latent Markov chain and the observations are independent conditionally on the states. Such models are very flexible since the determination of the states is data-driven instead of depending on arbitrarily chosen exogenous variables. Furthermore, they allow nonparametric state-dependent distributions, and then, using a few parameters, they can model complex time and space dependence. We may recall that it is not surprising from probabilistic point-of-view to use latent variables for modeling complex dependencies, since we know that general exchangeable random variables can be realized as a mixture of product distributions (for some latent distribution) thanks to the De Finetti theorem [Diaconis and Freedman, 1980]; in our setting, exchangeability is not a priori satisfied but still, this case is still inspiring from modeling dependencies.

Time-homogeneous HMMs are generally used for multisite generation, either of rainfall occurrences [Zucchini and Guttorp, 1991] or of the whole rainfall field. In [Kirshner, 2005], the author has proposed an overview and tests different options for multivariate emissions, from conditional independence to complex dependence structures, going through tree structures. The authors of [Ailliot et al., 2015a] have offered a more recent overview of weather type-based stochastic weather generators, including HMMs. Because the behavior of weather variables depends much on the season or on the time of the day, it is easy to believe that allowing non-homogeneity of the HMM likely gives better results. Therefore, extensions to time-nonhomogeneous HMMs have also been proposed, for instance to introduce a diurnal cycle [Ailliot and Monbet, 2012, Ailliot et al., 2015b]. Combining HMM and weather types variables is also possible, see for instance [Hughes and Guttorp, 1994a, Hughes and Guttorp, 1994b, Hughes et al., 1999].

In [Touron, 2019a] the author has designed a multivariate (temperature and precipitation) single-site weather generator based on nonhomogeneous HMM to take the seasonality as well as possible trends in climate variables into account. Focusing on precipitation, different models use the nonhomogeneous HMM [Hughes and Guttorp, 1994a, Hughes and Guttorp, 1994b, Hughes et al., 1999, Bellone et al., 2000]. [Holsclaw et al., 2016] used a Bayesian HMM for the climate downscaling of multisite precipitation in South and East Asia, while [Kroiz et al., 2020] proposed a daily precipitation generator based on HMM

to study the Potomac River Basin in Eastern USA over the wet season months. In this last study, Gaussian Copula are used to improve the spatial correlation.

Our contributions. We present a seasonal model based on a Hidden Markov Model that we name Seasonal Hierarchical Hidden Markov Model (SHHMM) to generate spatio-temporally realistic weather series, here precipitation. As in [Touron, 2019b], we consider a fully nonhomogeneous model where all parameters change periodically within the year. However, it is multisite as in [Kroiz et al., 2020], hence, the hidden states will be used to reproduce spatial patterns. In order to do that correctly, rather than having continuous emission distribution (rain amount) in the HMM like [Kroiz et al., 2020], we focus on a discrete one i.e., rain occurrence, as in Bernoulli mixture. Training hidden state models with binary variable such as wet/dry is well established in Machine Learning classification techniques, see [Bishop, 2006]. Discrete emission distributions might sound like a simplification compared to what is done in the literature where the rain amount R is directly express as a mixture of an atomic and continuous distribution [Touron, 2019a, Kroiz et al., 2020, Holsclaw et al., 2016], or using censored Gaussian [Ailliot et al., 2009, Baxevani and Lennartsson, 2015] or in the context of Markov Switching Models [Ailliot and Monbet, 2012, Ailliot et al., 2015b, Monbet and Ailliot, 2017, Ailliot et al., 2020]. However, we show in this paper how our approach produces fully interpretable hidden states relevant not only for rain occurrence. This is made possible by our assumption described in Section 2.2 that forces the hidden states to learn correlations. More complex HMM are also proposed, trained on multivariate weather variables with the risk of being increasingly difficult to train, and with loss of interpretability. See [Pohle et al., 2017, de Chaumaray et al., 2023] for discussions on how imperfect parametric emission distributions can influence, e.g., overestimate the number of hidden states. For example, extreme precipitations are often outside the reach of standard parametric rain distributions, here these types of events could disproportionally assign the weather regimes.

Moreover, the discrete/continuous nature of precipitation is also not well suited for the inclusion of previous day history. To circumvent this issue and correctly reproduce the distribution of dry and wet spells, our model includes (see Section 2.3) an additional hierarchical Markov dependence as describe in [Cappé et al., 2005, Section 2.2.3], also called Auto Regressive HMM [Kirshner, 2005, Section 3.1.1] by some authors.

Finally, our model also includes the rainfall amounts, which are added to the SHHMM and conditionally to the hidden states. In this case, this hidden states act as exogenous variables as in [Vrac et al., 2007] for the rain amount variables.

The combination of interpretability, Bernoulli mixture, plus nonhomogeneous hierarchical Markov dependence and modular structure where extra layers benefit from the hidden state as exogenous variable is new. Moreover, it performs very well in terms of dry and wet spells distributions, even extreme ones. These properties are indeed important for the desired application to the study of extreme dry sequences and their possible change in the future due to climate change in the context of hydropower or nuclear power generation. Ultimately, we show how to use our model with climate change scenarios and analyze the change in terms of parameters and extremes.

The model and its code are available in the Julia package `StochasticWeatherGenerators.jl` [Métivier, 2024]. It contains a reproducible step-by-step tutorial in its documentation describing all the data loading, training process and simulations of the model describe in this paper. Most figures of

this paper are exactly reproduced using the tutorial.

Organization of the paper. The paper is designed to offer an incremental construction of the model, with validation as it goes along. In Section 2, we describe step-by step the construction of the SHHMM model. We explain in Section 3 the procedure to infer and select the model. The Section 4 is entirely dedicated to the interpretation of the model parameters, in particular the trained hidden states are interpreted as Global Weather Regimes for France and will be compared to the North Atlantic Oscillations. In Section 5 we show simulations results for the spatio-temporal rain occurrence sequences with a special focus on extreme dry/wet sequences. The actual rain amounts are then added on top of the previous model in Section 6 and then tested in simulations in Section 7. In Section 8, we train our model on a climate change scenario and discuss the results.

Notations used in the paper. For a positive integer M, we set $[\![1 : M]\!] := \{1, 2, \ldots, M-1, M\}$. If Θ is a finite set, $|\Theta|$ denotes its cardinality. We make the distinction between t for a day and n for a date, see Subsection 2.2. The number of days is $T = 366$.

2 Hidden Markov Chain Modeling

In this section, we design the various statistical models studied in the work.

2.1 Data

The data is extracted from the European Climate Assessment & Dataset website [ECAD, 2022]. We focus only on stations in France. Among the 72 French (+2 from Belgium and one from Luxembourg) weather stations, 31 stations have 100% valid data from Jan. 1st 1956 to Dec. 31st 2019, i.e., a 64-year range and, 23376 data rows. We select $S = 10$ of these stations, well spread in all of France: these weather stations are indexed with $s \in \mathcal{S} := \{1, \cdots, S\}$. The station repartition will be justified in Section 2.2 where the conditional independence hypothesis is presented. We show on Figure 1 all the selected stations; in addition, we report in the heatmap scale of the historical maximum of consecutive days without rain – dry spell – at each location. One of the goals of our modeling is to reproduce similar records. Only in Section 8, we will investigate how our model (and its parameters) evolves when historical data are replaced by future projection data according to some RCP scenarios.

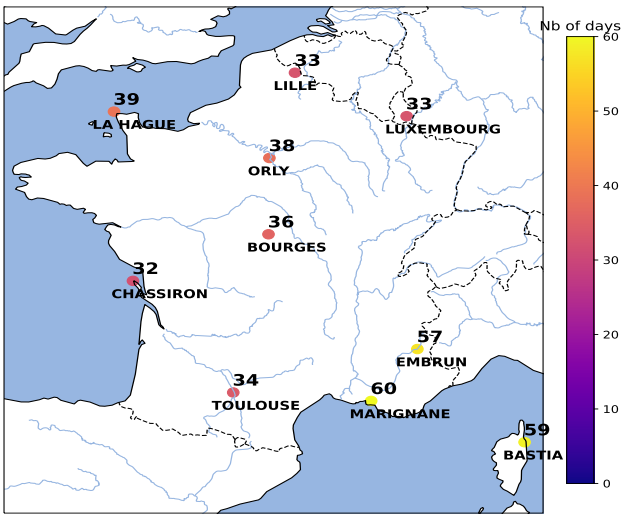


Figure 1: The 10 selected stations with their respective dry spell historical records over the period 1956-2019.

The $N = 23376$ consecutive weather observations are labeled with $n \in \mathcal{D} := [\![1, N]\!]$. The first weather observable under study is the Multisite Rain Occurrence (MRO in short)

$$Y^{(n)} := (Y_1^{(n)}, \ldots, Y_S^{(n)}) \in \mathcal{I} := \mathcal{I}_s^S$$

at date n, where each $Y_s^{(n)} \in \mathcal{I}_s := \{\text{dry}, \text{wet}\}$ where dry means no rain and wet means nonzero rain, i.e., above 0.1 mm of daily cumulated rain.

The next subsections are devoted to the design of the model for the evolution of the MRO. The actual nonzero rain amount will be added on top of the model after it is trained in Section 6. Our approach relies on Hidden Markov Model: generally speaking, it is made of a hidden component $\{Z^{(n)} : n \geq 1\}$ (that should be inferred) and of an observed one $\{Y^{(n)} : n \geq 1\}$ (here the MRO). All processes are discrete-time processes. See [Cappé et al., 2005] for a general account about Hidden Markov Models.

2.2 Seasonal Hidden Markov Model (SHMM), model $\mathcal{C}_0$

For the sake of clarity, we start with a simplified model, which will be extended hereafter. Consider first the hidden component Z, common to all stations $s \in \mathcal{S}$: it can take discrete values in $\mathcal{K} := [\![1 : K]\!]$ which will be later interpreted as climate states for the region of interest, here France. We will thus refer to this variable as a Global Weather variable, as often done in the literature [Holsclaw et al., 2016, Kroiz et al., 2020]. The time-evolution of $\{Z^{(n)} : n \geq 1\}$ follows a nonhomogeneous Markov Chain on the state space $\mathcal{K}$, with initial distribution $\xi = (\xi_1, \cdots, \xi_K)$, i.e. $\xi_k = \mathbb{P}\left(Z^{(1)} = k\right)$, and transition matrix $Q_n \in \mathbb{R}^{K \times K}$ for $n \geq 1$,

$$Q_n(k, k') = \mathbb{P}\left(Z_{n+1} = k' \mid Z_n = k\right).$$

To fit the climate context, we assume that the transition matrix Q_n is a T-periodic function of n with $T = 366$, i.e. $Q_{n+T} = Q_n$; we will thus refer to the Markov Chain as *Seasonal* Markov Chain. In that case, we will distinguish between the label *day of the year* $t \in \mathcal{T} := [\![1 : T]\!]$ and the label *full date* n used to denote the position in the sequence. At each n corresponds one t, but for each t there are as many n as the number of periods in the sequence: the matrices Q depend on time only through the day t.

Next, we design the model for the time-evolution of the MRO Y. The intuition behind the choice of well spread stations is that local weather variables Y conditionally to global weather variables Z are independent. In addition, we assume that the conditional distribution of $Y^{(n)}$ does not depend on the past of Y and is also periodic. All is summarized in the following assumption.

(H-$\mathcal{C}_0$) Z evolves as Seasonal Markov Chain with period $T = 366$. Conditionally to the process $\{Z^{(n)} : n \geq 1\}$, the spatial components $Y_1^{(n)}, \ldots, Y_S^{(n)}$ are independent and, furthermore, the conditional distribution of each $Y_s^{(n)}$ only depends on $Z^{(n)}$. This is a Bernoulli distribution describing the probability of rain at station s and date n, conditionally to $Z^{(n)} = k$: it is denoted $f_{k,n,s}$ (called *emission distribution*) and assumed T-periodic, i.e. $f_{k,n+T,s} = f_{k,n,s}$, and thus represented as

$$f_{k,t,s}(y_s) = \mathbb{P}\left(Y_s^{(n)} = y_s \mid Z^{(n)} = k\right) = \lambda_{k,t,s}\mathbf{1}_{y_s=w} + (1 - \lambda_{k,t,s})\mathbf{1}_{y_s=d} \tag{1}$$

for some parameters $\lambda_{k,t,s} \in [0, 1]$.

The above model for $\left\{\left(Z^{(n)}, Y^{(n)}\right) : n \geq 1\right\}$ is referred[1] to $\mathcal{C}_0$ and called Seasonal Hidden Markov Model (SHMM), with period T, initial distribution $\xi_\cdot = \mathbb{P}\left(Z^{(1)} = \cdot\right)$, transition matrix Q_t, and emission distributions $f_{k,t,s}$. This SHMM terminology is borrowed to [Touron, 2019a].

If T was equal to 1, this SHMM would be a regular homogeneous HMM. The SHMM chain is illustrated on Figure 2.

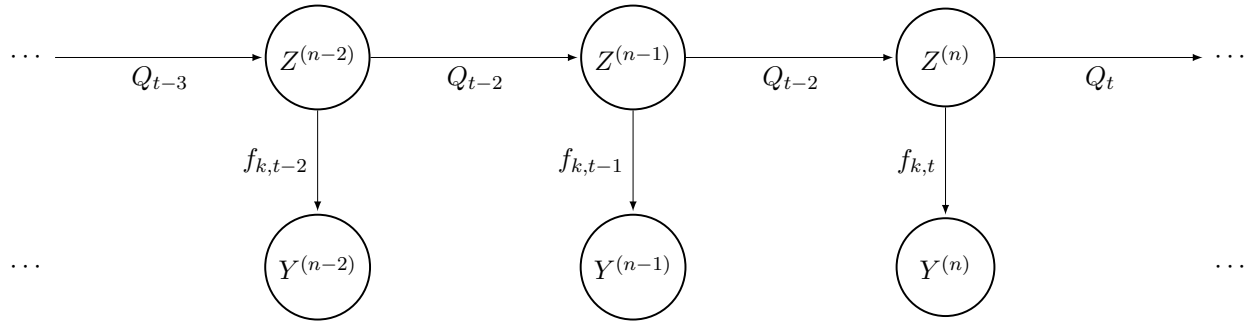


Figure 2: A Seasonal Hidden Markov Process $(\xi, Q_t, f_{k,t})_{k\in\mathcal{K},t\in\mathcal{T}}$

A few remarks before going further.

- This model accounts for leap years: for instance, the date $n = 59+366 = 425$ corresponds to Feb. 28, 1957, i.e., to day $t = 59$ while the next date $n = 426$, March 1, 1957, is day $t = 61$. All Feb 29 are labeled with $t = 60$. With this convention, the estimation of parameters for $t = 60$ will be

[1] the index $m = 0$ in $\mathcal{C}_0$ referring to the no-dependence of Y in its past

performed with 3 times less data than for other dates: nevertheless, it will have a quite minor impact on the procedure because of the time-smoothing of parameters exposed in Subsection 2.4.

- The annual periodicity of the distributions Q_t and $f_{t,k,s}$ is questionable. On the one hand, for obvious reasons of statistical inference, it is not possible to try to estimate as many distributions (parameterized by $n \in \mathcal{D}$) as there are data available, which leads to the reasonable assumption of annual stationarity as in [Touron, 2019b]. On the other hand, annual stationarity is probably not accurate considering climate change. In our methodology, the calibrated parameters should be understood as valid over the data horizon used. We will see in Section 8 that shifting the data period into the future (using climate projection under different RCPs) will cause some parameters to evolve. Let us mention some tests in [Touron, 2019b, Chapter V] showing that the effect of climate change on precipitation is not easily identifiable (unlike for temperatures), supporting the stationarity hypothesis of our model.

As a consequence of the spatial independence assumption in **(H-$\mathcal{C}_0$)**, the conditional likelihood of the MRO at date n is given

$$f_{k,t}(y) := \mathbb{P}\left(Y^{(n)} = y \mid Z^{(n)} = k\right) = \prod_{s \in \mathcal{S}} f_{k,t,s}(y_s). \tag{2}$$

This probability depends on n only through the day t. This assumption forces the model to learn spatial features (and spatial dependence) through the hidden states.

Later in this paper, we show (see Figure 23) that this SHMM produces in general shorter dry or wet spells than the ones observed, suggesting that the Markovian dynamics of the Global Weather Variable Z is not enough to stochastically explain the temporal evolution of the MRO Y. Indeed, Z is a Global Weather Variable over all France and does not take into account the local dynamics of rain occurrence $Y_s^{(n)}$, i.e. that in addition to being influenced by the global weather, local weather should also be dependent on the local previous day's MRO $Y^{(n-1)}, Y^{(n-2)}, \cdots$. Hence, it makes sense to define the dynamics of the MRO conditionally to several previous days. This is the *raison-d'être* of the next models $\mathcal{C}_m$, $m > 0$.

We end this section with another set of remarks considering our model assumptions, that will also apply for $\mathcal{C}_{m>0}$ models.

- The conditional independence hypothesis (2.2) is a fundamental hypothesis for our model. Indeed, as described in the introduction, this choice forces the model to learn spatial correlation exclusively through the hidden state. Allowing conditional dependence, e.g., as in [Zucchini and Guttorp, 1991, Eq. (22)], increases dramatically the number of parameters to fit and can lead in an extreme case to all correlations being learned by the conditional probability with irrelevant hidden states. In practice, this hypothesis can be checked a posteriori, see Figure 13 and is valid for stations being far enough apart.

- The choice of Bernoulli mixtures as emission distribution for MRO and then the addition in Section 6 of rain amount might look like an unnecessary two steps model. Indeed, traditionally, models directly consider emission distribution of rain amount at once. In this case, as suggested

in [Pohle et al., 2017] the parametric choice of the rain distribution might strongly influence the model selection by overestimating the number of hidden state K. The choice of Bernoulli distributions for binary variables is however exact, suggesting that our model will likely pick a smaller number of hidden states, i.e., more interpretable.

2.3 Seasonal Hierarchical Hidden Markov Model (SHHMM), model $\mathcal{C}_m$ with $m > 0$

To reproduce better the dry and wet spell distributions, we consider additional local conditioning. Different length of this additional local conditioning $Y_s^{(n)} \mid (Z^{(n)}, Y_s^{(n-1)}, Y_s^{(n-2)}, \cdots, Y_s^{(n-m)})$ will correspond to different models $\mathcal{C}_m$ (with some *memory parameter* $m = 1, 2, \cdots$). Intuitively, models with history $\mathcal{C}_{m>0}$ should display better temporal persistence than the $\mathcal{C}_0$ model, i.e., consecutive day sequence statistics should be replicated better. On the other hand, these models $\mathcal{C}_{m>0}$ require more parameters to be fitted for the same number of data, thus one should expect statistically less accurate estimates if m is too large.

Given $m > 0$, we introduce the history variable

$$H^{(n)} := (Y^{(n-1)}, Y^{(n-2)}, \cdots, Y^{(n-m)}) \in \mathcal{H}^{(m)} := \mathcal{I}^m$$

and its local analog $H_s^{(n)} := (Y_s^{(n-1)}, Y_s^{(n-2)}, \cdots, Y_s^{(n-m)}) \in \mathcal{H}_s^{(m)} := \mathcal{I}_s^m$. The following hypothesis $\mathcal{C}_m$ summarizes the model.

(H-$\mathcal{C}_m$) Z evolves as Seasonal Markov Chain with period $T = 366$. Conditionally to the Global Weather process $\{Z^{(n')} : n' \geq 1\}$ and to the history of local weather $H^{(n)}$, the spatial components $Y_1^{(n)}, \ldots, Y_S^{(n)}$ are independent and, furthermore, the conditional distribution of each $Y_s^{(n)}$ only depends on $Z^{(n)}$ and $H_s^{(n)}$. This emission distribution is a Bernoulli distribution describing the probability of rain at station s and date n, conditionally to $Z^{(n)} = k$ and $H_s^{(n)} = h_s$: it is denoted f_{k,n,s,h_s}, assumed T-periodic, and represented as

$$f_{k,t,s,h_s}(y_s) := \mathbb{P}\left(Y_s^{(n)} = y_s \mid Z^{(n)} = k, H_s^{(n)} = h_s\right) = \lambda_{k,t,s,h_s} \mathbf{1}_{y_s=w} + (1 - \lambda_{k,t,s,h_s}) \mathbf{1}_{y_s=d}, \quad (3)$$

for some parameters $\lambda_{k,t,s,h_s} \in [0, 1]$ depending on n only through the day t, the hidden state k and on the m previous day observations value h_s at station s.

As a consequence, and similarly to (2),

$$f_{k,t,h}(y) := \mathbb{P}\left(Y^{(n)} = y \mid Z^{(n)} = k, H^{(n)} = h\right) = \prod_{s \in \mathcal{S}} f_{k,t,s,h_s}(y_s) \quad (4)$$

The $\mathcal{C}_{m>0}$ models are defined like SHMM by $(\xi, Q_t, f_{k,t,h})_{k \in \mathcal{K}, t \in \mathcal{T}, h \in \mathcal{H}^{(m)}}$ where the law of the first observations $\xi_\cdot = \mathbb{P}\left(H^{(1)} = \cdot, Z^{(1)} = \cdot\right)$ where $H^{(1)} = (Y^{(0)}, \ldots, Y^{(1-m)})$ is added.

Regarding the usual terminology of Hidden Markov Chain, the model $\mathcal{C}_0$ is a standard (periodic) HMM ([Cappé et al., 2005, Section 2.2]) since the observed variables $\{Y^{(n)} : n \geq 0\}$ are independent conditionally on the hidden variables $\{Z^{(n)} : n \geq 0\}$. For other models $\mathcal{C}_1$, $\mathcal{C}_2$, $\cdots$, because of the dependence with respect to previous days through $Y^{(n-1)}, Y^{(n-2)}, \cdots$, we are rather in the presence of Hierarchical HMMs as described in [Cappé et al., 2005, Section 2.2.3] (also named Auto Regressive

HMM [Kirshner, 2005, Section 3.1.1]): conditionally to $\{Z^{(n)} : n \geq 0\}$, the MRO process $\{Y^{(n)} : n \geq 0\}$ evolves as a Markov chain with memory m.

This is a significant difference with other precipitation models in the literature, like [Touron, 2019a, Holsclaw et al., 2016, Kroiz et al., 2020] and we will use the denomination SHHMM for Seasonal Hierarchical Hidden Markov Model to denote $\mathcal{C}_{m>0}$. We illustrate a $m = 1$ day memory SHHMM in Figure 3.

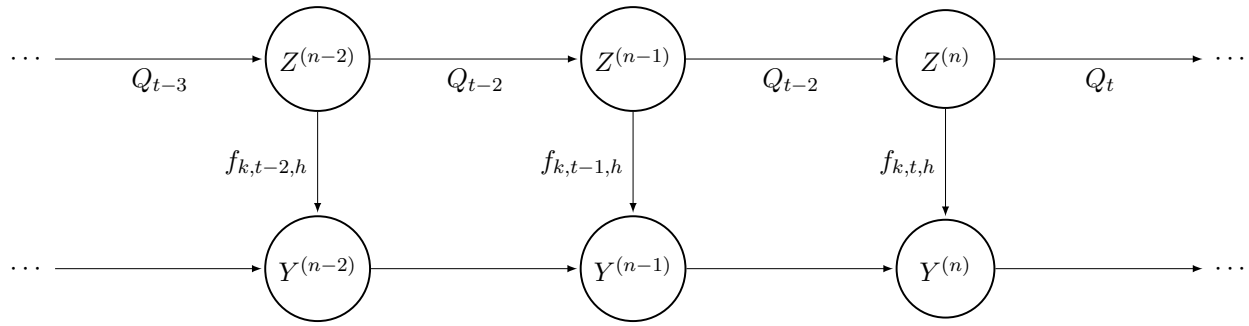


Figure 3: Illustration of a Seasonal Hierarchical Hidden Markov Model with one-day memory $m = 1$.

Note that some authors have already considered similar (time independent) hierarchical HMM models, e.g., [Kirshner, 2005, Section 6.1.1]. However, the combination of discrete (Bernoulli) emissions, seasonality (see Section 2.4) and interpretability (see Section 4) is something new as far as we know.

2.4 Hypothesis and modelling of the time-regularity of parameters

The previous models $\mathcal{C}_0$ and $\mathcal{C}_{m>0}$ depend on the T-periodic functions $(Q_t, f_{k,t,h})_{k\in\mathcal{K},h\in\mathcal{H}^{(m)}}$. A quick inspection of the number of scalar parameters to estimate at each day $t \in \mathcal{T}$ gives

- $K(K-1)$ coefficients for the transition matrix Q_t,
- $K \times S \times 2^m$ coefficients for Bernoulli distribution parameters λ_{k,t,s,h_s} for all k, s, h_s.

For $S = 10$, $K = 4$, $m = 1$, it gives 92 scalar parameters, which is larger than the number of available data at each day t (64 for usual days and 16 for the Feb 29). On the one hand, estimating the parameters by maximizing the observed likelihood independently at each day $t \in \mathcal{T}$ is conceptually simple. On the other hand, the estimated parameters would suffer from a high variance as there are too few data at each day t. Therefore, in the inference procedure that will be exposed in Section 3, a time-regularity constrain will be imposed. This procedure (detailed later) will be essential to recover interpretable and meaningful results.

Let us argue in more details. Intuitively, the timescale of variation of the model parameters should be in the order of magnitude of a month (30 days). Hence, once fitted, the parameters should evolve as a smooth function of day t. The advantages of imposing a smoothing are multiple:

(1) This corresponds to the physical intuition.

(2) It helps to overcome the lack of data at each day t, indeed time-regularization implies that the data from neighboring days $t-1, t+1, t-2, t+2, \dots$ are somehow accounted when making inference at day t.

(3) It can help with model selection of weather regime number K: very large K models will overfit the data, exhibiting apparently very good performance. However, once calibrated the time-regularity is poor (jumps...), which is against the physical intuition.

(4) In terms of identifiability of the model, it is well known that HMM are identifiable up to relabelling of the hidden states. In the case of SHMM, the model is not identifiable up to the relabeling of hidden states at each day t [Touron, 2019a]. Thus, it is very likely that a naive likelihood optimization routine gives quite different parameters on consecutive days, whereas for obvious interpretability reasons, we seek a smooth evolution as a function of the day t of the calendar year.

In practice, there are two possible approaches to account for time-regularity of the estimated parameters:

- Optimize then Smooth: Once all the parameters for all days are estimated independently, after some adhoc labeling procedure of the hidden states on each day t, perform a time-denoising procedure (kernel method, least-squares regression, ...) with some meta-parameters.
- Smooth then Optimize: assume some parametric form to parameters as a function of the day $t \in \mathcal{T}$, see Eqs. (5)-(6) below, and directly infer new parameters without worrying about smoothness and day-to-day relabeling. The final SHMM or SHHMM is then only identifiable up to a global relabeling common to all t.

According to some tests that are not reported here, we have observed that the second approach gives much better results in terms of regularity and variance and is much more robust. We adopt this approach from now on. Thus, each parameter $(Q_t, \lambda_{k,t,s,h_s})_{k\in\mathcal{K},s\in\mathcal{S},h\in\mathcal{I}_s}$ is composed with the trigonometric polynomial as follows: given some coefficients $c_0, c_1, \dots,$ set

$$P_c(t) := c_0 + \sum_{j=1}^{\mathrm{Deg}} \left(c_{2j-1} \cos\left(\frac{2\pi}{T} jt\right) + c_{2j} \sin\left(\frac{2\pi}{T} jt\right)\right), \tag{5}$$

for some degree Deg. For all $k \in \mathcal{K}$ the transition matrices are given by

$$Q_t(k,l) = \frac{e^{P_{c_{k,l}}(t)}}{1+\sum_{l=1}^{K-1} e^{P_{c_{k,l}}(t)}} \quad \text{for} \quad 1 \le l < K, \qquad\qquad Q_t(k,K) = \frac{1}{1+\sum_{l=1}^{K-1} e^{P_{c_{k,l}}(t)}},$$

and the Bernoulli parameters in (1)-(3) by

$$\lambda_{k,t,s,h_s} = \frac{1}{1+e^{P_{c_{k,s,h_s}}(t)}}.$$

The parametrization of Q_t corresponds to the log-ratio transformation well-known in Compositional Data Analysis [Pawlowsky-Glahn and Buccianti, 2011]. These definitions ensure $0 < \lambda_{k,t,s,h_s} < 1$,

$0 < Q_t < 1$ and $\sum_{l\in\mathcal{K}} Q_t(k,l) = 1$, $\forall t \in \mathcal{T}$ and $\forall k \in \mathcal{K}$. A model with high degree Deg will be able to capture shorter and shorter sub-seasonal/monthly/sub-monthly/etc. phenomena.

A quick inspection of the number of parameters (the coefficients c) gives (for $S = 10$, $K = 4$, $m = 1$, and Deg $= 2$ corresponding to roughly 4 seasons) $92 \times 5 = 460$ scalar parameters (for all $t \in \mathcal{T}$) instead of $92 \times 366 = 33672$ in the previous day-by-day parametrization. The gain is quite significant. However, the maximization step has no analytical solutions: the subsequent numerical optimization is heavy due to the fact that now, $(Q_t, \lambda_{k,t,s,h_s})_{k\in\mathcal{K},s\in\mathcal{S},h\in\mathcal{I}_s}$ depends on each other for different $t \in \mathcal{T}$. The resulting parametric problem is of lower dimensions, but more complex to solve than the T individual problems.

In the sequel, we denote by θ all the coefficients appearing in (5)-(6) and that are to be optimized:

$$\theta := \{c_{k,l} \in \mathbb{R}^{2\text{Deg}+1}, c_{k,s,h_s} \in \mathbb{R}^{2\text{Deg}+1} : k \in \mathcal{K}, l \in [\![1 : K-1]\!], s \in \mathcal{S}, h_s \in \mathcal{I}_s\}. \tag{7}$$

2.5 Identifiability

For the inference problem to make sense, the model must be identifiable. Latent models are known to be only identifiable up to label swapping. Moreover, Bernoulli mixtures are known to be non-identifiable [Gyllenberg et al., 1994]. However, they are identifiable under a weaker notion of *generic identifiability* up to label swapping if the following condition holds [Allman et al., 2009, Corollary 5]

$$2\lceil \log_2(K) \rceil + 1 \leq S. \tag{8}$$

Generically identifiable [Allman et al., 2009] implies in particular that the set of points for which identifiability does not hold has measure zero. Hence, for the applications, this notion is enough. For our application, we explore K being at most 8 so that $S \geq 7$.

In [Touron, 2019a, Theorem 1], the identifiability up to label swapping of Seasonal Hidden Markov Model is proven under the following assumptions

(1) For $1 \leq t \leq T$, the transition matrix Q_t^* are invertible and irreducible

(2) The matrix $Q_1^* \dots Q_T^*$ is ergodic, and its unique stationary distribution ξ^* is the distribution of Z_1

(3) For each $t \in [\![1, T]\!]$, the K emission distributions $(\nu_k^*(t))_{k\in\{1,\dots,K\}}$ are linearly independent.

The star $*$ denote the set of true parameter. The irreducibility and ergodicity are satisfied under the parametric assumption for Q_t since all the matrix coefficients are strictly positive. The invertibility of the Q_t is proved to hold up to a negligible set of parameter [Touron, 2019a, Section 2.4.1] for our parametric choice. The second condition can be shown using that the coefficients of Q_t are strictly positive, so those of $Q_1^* \dots Q_T^*$ also, therefore $Q_1^* \dots Q_T^*$ is irreducible and aperiodic. To prove that the third assumption is satisfied in our case, we use the equivalence [Yakowitz and Spragins, 1968, Theorem Section 3] between linear independence of K emission distributions $(\nu_k)_{k\in\mathcal{K}}$ and the identifiability of the mixture $\sum w_k \nu_{k,t}^*$ for some weights $(w_k)_{k\in\mathcal{K}}$. Together with the condition Eq. (8), it follows that the model $\mathcal{C}_{m=0}$ is generically identifiable up to a global relabeling. For higher order models $\mathcal{C}_{m>0}$, the local memory (autoregressive properties) of the emission distribution prevents from directly applying the previous results, however one can expect similar condition to holds true.

3 Fitting the SHHMM and selecting the hyper-parameters

In [Touron, 2019a], the authors also prove that the maximum likelihood estimator is a consistent estimator for the Seasonal HMM, i.e. $\mathcal{C}_{m=0}$. Proving the consistency for the hierarchical model $\mathcal{C}_{m>0}$ is outside the scope of this paper, however we will still use the maximum likelihood estimator to infer the model parameters.

Maximizing the likelihood of a latent model is usually done with Expectation Maximization (EM) algorithm. See [McLachlan and Krishnan, 2007] for a general review on the EM algorithm and its extensions. To maximize the loglikelihood of the SHHMM, we will use a heterogeneous version of the Baum-Welch algorithm, which is a special kind of **EM** algorithm for Hidden Markov Models. The detail of the algorithm can be found in the Appendix D. A known issue of EM algorithms is that they can converge to local maxima. As we will illustrate, a naive random initialization of the algorithm without a good guess will likely either land in some meaningless local maxima – even if multiple random initial conditions are tried– and/or to take a very long time to converge.

Hence, before fitting SHHMM with the Baum-Welch algorithm, we will first find a crude estimator of the SHHMM by solving many simpler subproblems by the procedure described below.

3.1 Initialization: The Slice Estimate

The idea is to first treat the MRO observations of each day of the year $t \in \mathcal{T}$ separately. This procedure is in fact similar to the "Optimize then Smooth" procedure described in Section 2.4. On each day t, the emissions distributions $\{\widetilde{f}_{1,t}, \cdots, \widetilde{f}_{K,t}\}$ form a mixture model that can be fitted with a standard **EM** algorithm.

Once this is done, we relabel the hidden state at each day t to ensure some continuity in the estimated parameters $\widetilde{\theta}_{k,t,h,s}$. Finally, by identifying the most likely *a posteriori* states on each date n, we obtain an estimated sequence, $\{\widetilde{z}^{(n)} : n \in \mathcal{D}\}$ which we use to fit the transition matrices $\hat{Q}(t)$. The whole procedure is described in Appendix E.

3.2 Baum-Welch algorithm for SHHMM

In the previous section, we provided an estimate SHHMM that we will use as a starting point in the Baum-Welch algorithm. The algorithm alternates between Estimation (**E**) and Maximization (**M**) steps to converge to a local maximum of the observed likelihood defined for the SHHMM $(\xi, Q_t, f_{k,t,h})_{k\in\mathcal{K},t\in\mathcal{T},h\in\mathcal{H}^{(m)}}$ with $m \geq 1$ (see Section 2.3) by

$$\begin{aligned} L\left(y^{(1:N)};\theta\right) &= \mathbb{P}\left(Y^{(1:N)} = y^{(1:N)}\right) \\ &= \sum_{z_1,\ldots,z_N \in \mathcal{K}^N} \xi_{z_1,h_1} f_{z_1,t_1}\left(y^{(1)} \mid h^{(1)}\right) \prod_{n=2}^{N} Q_{t_n}(z_{n-1}, z_n) f_{z_n,t_n}\left(y^{(n)} \mid h^{(n)}\right), \end{aligned}$$

where for sake of simplicity we assume that h_1 is known, so that $\xi_{z_1,h_1} = \mathbb{P}\left(Z^{(1)} = z_1, H^{(1)} = h_1\right) = \mathbb{P}\left(Z^{(1)} = z_1\right)$. Note that this is the case in practice as we have a few extra data to define h_1. We briefly detail each steps of the EM algorithm, more details can be found in Appendix D.

- **Initialization.** An initial set of parameters $\theta^{(0)}$ is given initialization. As mentioned, we use for $\theta^{(0)}$ the parameters of the Slice Estimate SHHMM found in Section 3.1.

- **Step** $(i > 0)$,

  - **E-step**: compute the smoothing probabilities $\pi^{\theta^{(i)}}_{n|N}(k) = \mathbb{P}_{\theta^{(i)}}\left(Z^{(n)} = k \mid Y^{(1:N)}\right)$ and $\pi^{\theta^{(i)}}_{n,n+1|N}(k,l) = \mathbb{P}_{\theta^{(i)}}\left(Z^{(n)} = k, Z^{(n+1)} = l \mid Y^{(1:N)}\right)$ under the current parameter $\theta^{(i)}$. These probabilities can be computed using the Forward-Backward procedure described in Appendix D.

  - **M-step**: the function to maximize w.r.t. θ is

$$\begin{aligned} R\left(\theta, \theta^{(i)}\right) &= \mathbb{E}^{\theta^{(i)}}\left[\log L\left(Y^{(1:N)}, Z^{(1:N)}; \theta\right) \mid Y^{(1:N)}\right] \\ &= \sum_{k,l=1}^{K} \sum_{n=1}^{N-1} \pi^{\theta^{(i)}}_{n,n+1|n}(k,l) \log Q_{t_n}(k,l) + \sum_{k=1}^{K} \sum_{n=1}^{N} \pi^{\theta^{(i)}}_{n|N}(k) \log f_{k,t_n}(y^{(n)} \mid h^{(n)}) \\ &\quad + \sum_{k=1}^{K} \pi^{\theta^{(i)}}_{1|n}(k) \log \xi_k. \end{aligned}$$

- **Stop.** The iterations stop at $i = i_{\text{stop}}$ when $L(\hat{\theta}^{(i+1)}) - L(\hat{\theta}^{(i)}) < \epsilon_{\text{atol}}$.

Note that at **M-step**, the maximization can be done independently for the transition matrices and emission distributions (and initial distributions). However, since we enforce the coefficients $\theta^{(i)}$ to be periodic functions of the day of the year t, the maximization step cannot be done explicitly even for simple Bernoulli emission and is thus done numerically.

In all our numerical applications, the stopping criterion is $\epsilon_{\text{atol}} = 10^{-3}$. The loglikelihood at convergence is typically for the settings $K = 4$, $m = 1$, Deg $= 2$ and the historical data $L(\hat{\theta}^{(i_{\text{stop}})}) \simeq -116791$ i.e. $\epsilon_{\text{atol}}/|L(\hat{\theta}^{i_{\text{stop}}})| \sim 10^{-8}$. We also check that this stopping criteria is relevant for the θ parameters as we have $\max(|\theta^{(i_{\text{stop}})} - \theta^{(i_{\text{stop}}-1)}|) \simeq 10^{-3}$ where the max is taken as the largest difference between two iterations over all the parameters θ in Eq. (7).

To avoid being trapped in a local minimum, we run the algorithm 10 times with initial conditions randomized around the initial state $\theta^{(0)}$ provided in Section 3.1, see Appendix E.6 for more details. We then select the maximum likelihood amongst the different runs.

3.3 Hidden state inference: The Viterbi Algorithm

Once the SHHMM parameters are found $\hat{\theta}$, the most likely hidden states of the observed data sequence $\{\hat{z}^{(n)} : n \in \mathcal{D}\}$ can be inferred with the Viterbi algorithm [Viterbi, 1967]. The Viterbi algorithm is straightforward to adapt for periodic hidden chain.

3.4 Model Selection

We introduced three hyperparameters to our model: the local memory length $m = 0, 1, 2, \cdots$, the number of hidden states (weather regime) $K = 1, 2, 3, 4, \cdots$ and the degree $d = 0, 1, 2, \cdots$ of the trigonometric expansion Eq. (5). In particular, the number of hidden states K must be large enough

to reproduce spatial correlations but low enough to avoid overfitting and loss of interpretability. In principle, we could use different m_s at each station $s \in \mathcal{S}$, as well as different degree d for each type of variable and station (transition matrix coefficient, Bernoulli parameter, etc.). In this model, we fix m and d to be the same for all stations and variables.

In the literature several methods have been used to assess the best hyperparameters of more or less HMM, information criterion coefficients like the BIC, cross-validations see [de Chaumaray et al., 2023] and reference therein. From a theoretical point of view, no result guarantees the quality of these estimators for SHHMM. To select the hyperparameter K, we use the Integrated Complete-data Likelihood (ICL) criterion, as it favors non overlapping hidden states and show better good empirical performances with HMM [Celeux and Durand, 2008, Pohle et al., 2017]. It is defined as $L_C(y^{(1:N)}, z^{(1:N)}; \theta) = \mathbb{P}(Z^{(1:N)} = z^{(1:N)}, Y^{(1:N)} = y^{(1:N)}; \theta)$ which is in practice not accessible. The estimate $\hat{L}_C(y^{(1:N)}, \hat{z}^{(1:N)}; \hat{\theta}) = \mathbb{P}\left(Z^{(1:N)} = \hat{z}^{(1:N)}, Y^{(1:N)} = y^{(1:N)}; \hat{\theta}\right)$ uses the fitted parameter $\hat{\theta}$ and the decoded Viterbi most likely hidden state sequence $(\hat{z}^{(n)})_{n \in \mathcal{D}}$. The ICL is then computed as

$$\mathrm{ICL}(m, d, K) = \log(\hat{L}_C(y^{(1:N)}, \hat{z}^{(1:N)}; \hat{\theta})) - \frac{\log(N)}{2}|\hat{\theta}|.$$

The optimal $\{m, d, K\}$ set is obtained by maximizing $\mathrm{ICL}(m, d, K)$. In Figure 4, we see that $K = 4$, $m = 1$, $d = 2$ maximizes the ICL. Hence, for the rest of the paper unless specified otherwise, we will choose these parameters.

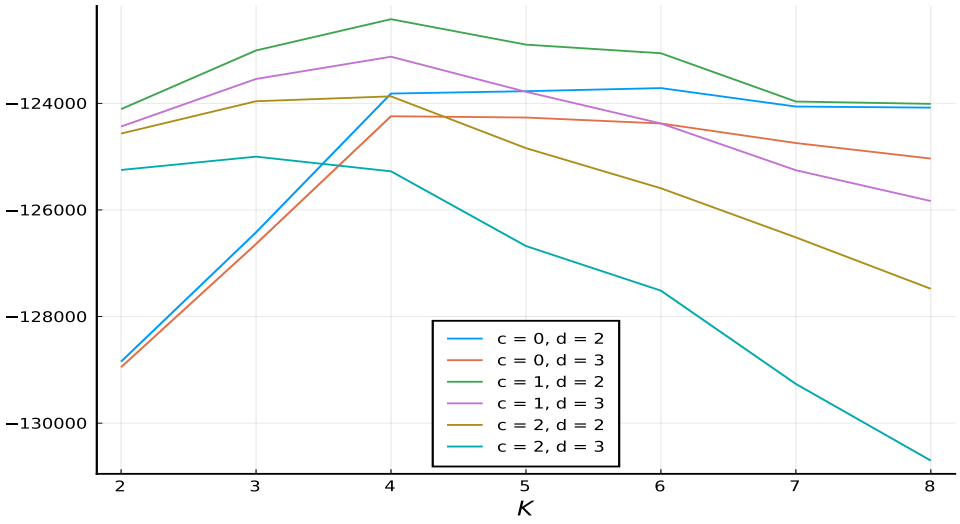


Figure 4: ICL for different values of the hyperparameters. The model $K = 4$, $m = 1$, $d = 2$ is the maximizer.

4 Interpretability: Making Sense of the Hidden States

One of the main messages of this paper is to show that the resulting hidden states are fully interpretable, both spatially and temporally. In particular, forcing conditional independence, see Eqs. (2), (4), forces

all spatial correlations to be in the hidden states. We describe in this section different points of view to give a sense of these hidden states that we also refer to as weather regimes.

In the following, all plots and interpretations are done for the model $\mathcal{C}_{m=1}$ with $K = 4$ and $d = 2$ which was the model selected in Section 3.4.

4.1 Spatial features

The hidden states have been introduced to give correlated rain events across France. Hence, we expect the hidden states to form some spatial patterns specific to the French weather, typically the south is generally dryer than the north.

Rain probability. In Figure 5, we show the rain probability given the hidden state k and that the previous day was dry, averaged for across the year.

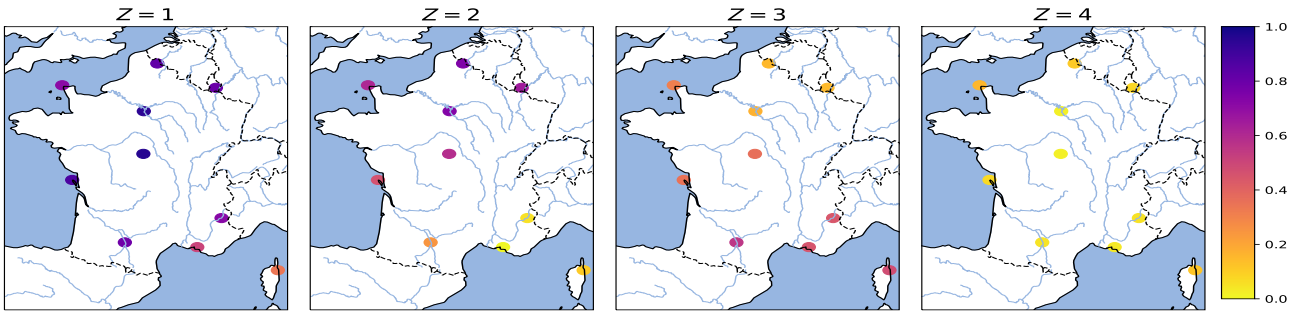


Figure 5: Yearly mean rain probability $T^{-1} \sum_{t \in \mathcal{T}} \lambda_{k,t,s,h}$ for $m = 1$ and $h =$ dry, i.e., the probability of rain at a location s, conditionally to the hidden state $Z = k \in [1, K = 4]$ and to a previous dry day.

The $Z = 1$ state corresponds to a high probability of rain over all France, $Z = 2$ correspond to a rainy climate in the north and drier in the south, $Z = 3$ is more or less the opposite, while in the state $Z = 4$ the probabilities of rain are low all over France. The trained model satisfyingly recovers known regional features of the French climate. For higher order models $K > 4$, the spatial features are more and more specific to peculiar scenarios, e.g., rainy only in Bastia. It can also be a signal of overfitting.

Pressure maps and North Atlantic Oscillations. In Figure 6, we show how the weather regimes are relevant in terms of pressure map.

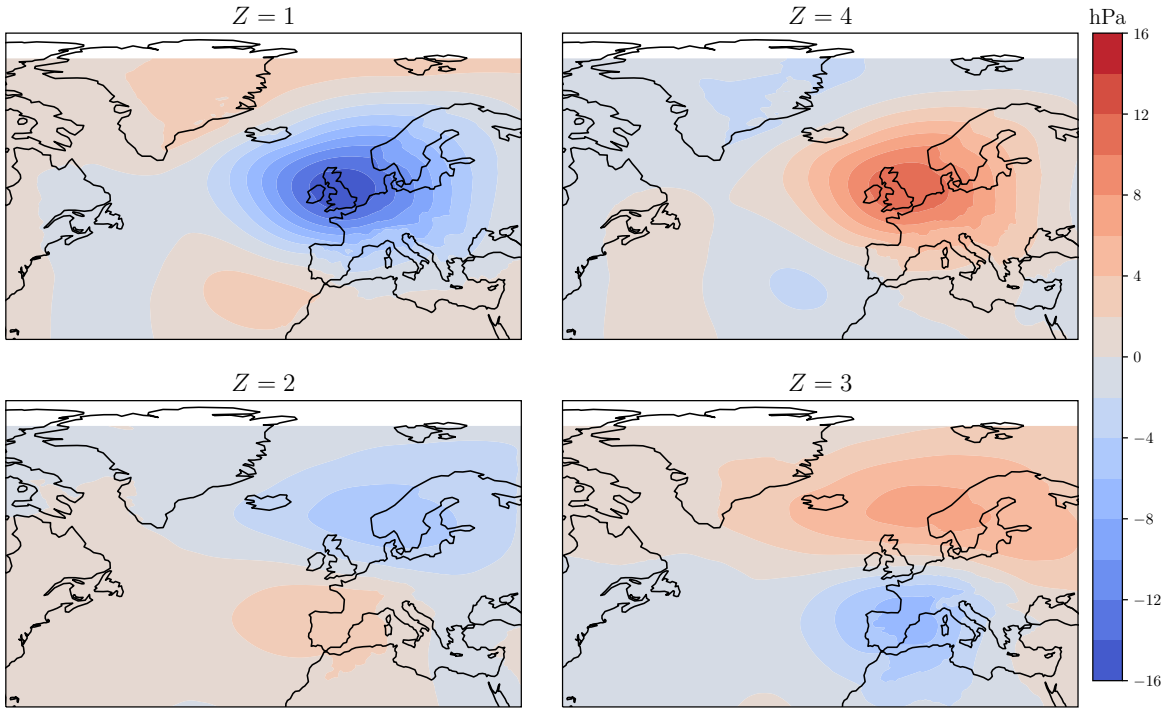


Figure 6: Winter (Dec-Jan-Feb) Mean Sea Level Pressure (Pa) anomalies (difference) between the average of all winter days in $Z = k$ state and the average of all winter days.

We consider the Mean Sea-Level pressure (MSP) from the Reanalysis ERA5 hourly data on single levels from 1979 to 2017 [Hersbach et al., 2020]. The pressure map is averaged over all winter days $\mathcal{D}_{\mathrm{W}} = \mathcal{D} \cap \{\text{December}, \text{January}, \text{February}\}$ conditionally to the hidden state (inferred before via the Viterbi algorithm, see Section 3.3) giving a pressure anomaly map $\Delta\text{MSE} = \mathbb{E}_{t\in\mathcal{D}_{\mathrm{W}}}(\text{MSP}(t) \mid \hat{z}_t = k) - \mathbb{E}_{t\in\mathcal{D}_{\mathrm{W}}}(\text{MSP}(t))$ at each longitude and latitude. The geographical area where the pressure maps are computed is (longitude $\in$ [80WEST, 40EST] and longitude $\in$ [25NORTH, 80NORTH]). It is much larger than France and corresponds roughly to the North Atlantic area. The results are shown on Figure 6.

The four maps clearly show four distinct regimes. These regimes can be compared with the well-known regime over the Europe-North Atlantic sector, among which the North Atlantic Oscillation (NAO). These are large scale weather regimes over the North Atlantic Ocean responsible for most of the climate variability [Woollings et al., 2010]. There are various definitions of the regimes of the North Atlantic weather based on the pressure at some northern and southern weather stations. For comparison, we have in mind the four Euro-atlantic regimes defined in [Cassou, 2004, Figure 7]. They display the same differential pressure map over winter months for a similar area. In [Cassou, 2004], the four regimes are NAO−, Atlantic Ridge, Blocking, NAO+. The two NAO regimes correspond to the reinforcement or attenuation of the Icelandic low and Azores high, leading to a strengthened or weakened westerly flow over France. The two other regimes correspond to different deviations of this flow, having different consequences for the French weather, depending on the season. In our model selection, we found the same number of hidden states $K = 4$, see Section 3.4. The order of magnitude

of the mean pressure anomalies between 0 Pa and 10 hPa is similar both in [Cassou, 2004, Figure 7] and Figure 6; the states defined by the SHHMM looks similar to the Euro-Atlantic regimes both in terms of order of magnitude and patterns.

- $Z = 1$ Rainy regime: depression all over France. It is similar to the NAO− state.
- $Z = 2$ Intermediate states rainier in the north: dipole of depression and anticyclone respectively in the north and south. It could be compared to the Atlantic Ridge state.
- $Z = 3$ Intermediate states rainier in the south: dipole of anticyclone and depression respectively in the north and south. It could be compared to the Blocking state.
- $Z = 4$ Dry regime: anticyclone all over France, similar to the NAO+ state.

It is remarkable that these large scale structures over Mean Sea Level Pressure are recovered with a model only trained over $S = 10$ stations all located in France with only dry or wet observations. Note that the large structure recovered resembles the traditional Euro-Atlantic regimes, but are more centered toward France due to the training data.

4.2 Seasonality

The SHHMM, transition matrix and emission distributions have periodic coefficients varying across the year. A consequence is that the hidden states are not fixed in time but can also vary. We expect variations to be smooth enough so that climate state $Z = k$ has a similar interpretation during the whole year.

4.2.1 Transition Matrix

We display, in Figure 7, the 16 coefficients of the transition matrix Q_t. The "dry" state $Z = 4$ is the

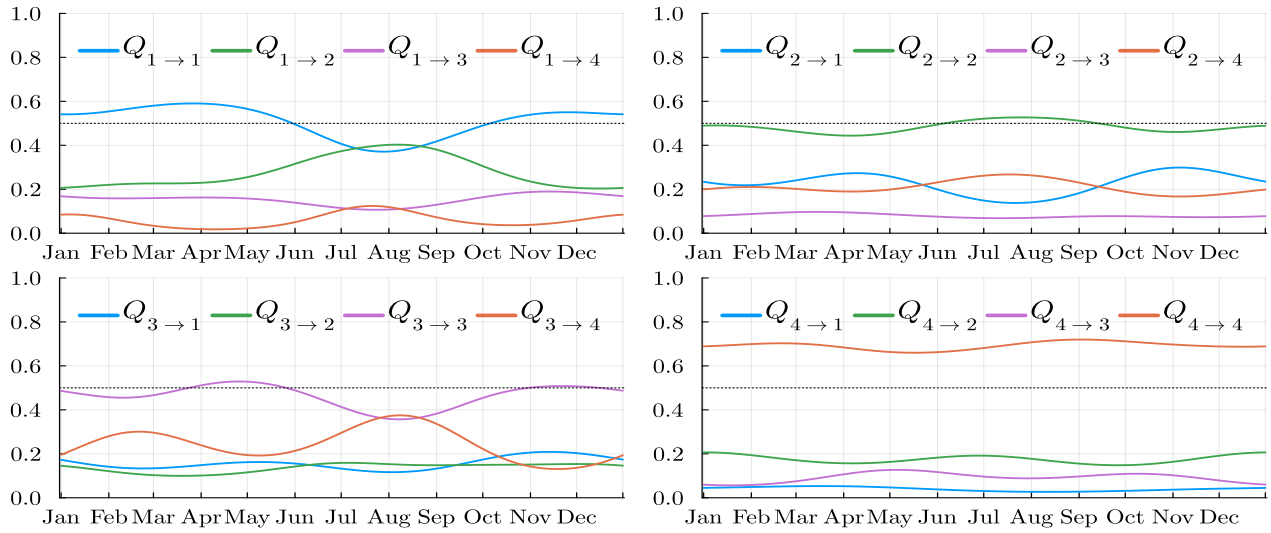


Figure 7: Temporal variation of the transition matrix $Q(t)$ for the SHHMM $K = 4$, $d = 2$ and $m = 1$.

state where the probability to stay in the same state is the highest. Hence, we expect longer global dry sequences than the other regimes. Probability to remain in the same state is the lowest in states 2 and 3, hence, these can be seen as transitional states. Moreover, state 4 has a very low probability to switch directly to state 1 (and vice versa) confirming that an intermediate state is required for this to happen. This makes sense with the intuition that a dry day all over the country is rarely followed by a wet day all over France. During some seasons, e.g., summer (Jun, Jul, Aug, Sep), state 2 will prefer to transition to a dry state 4 rather than the wet state 1. This is the opposite situation in the rest of the year. Again, this is consistent with the fact that during summer we expect the state 1 being less frequent.

4.2.2 Rain probability

We plot, in Figure 8, the rain probabilities in function of the station and climate variable $Z = k$.

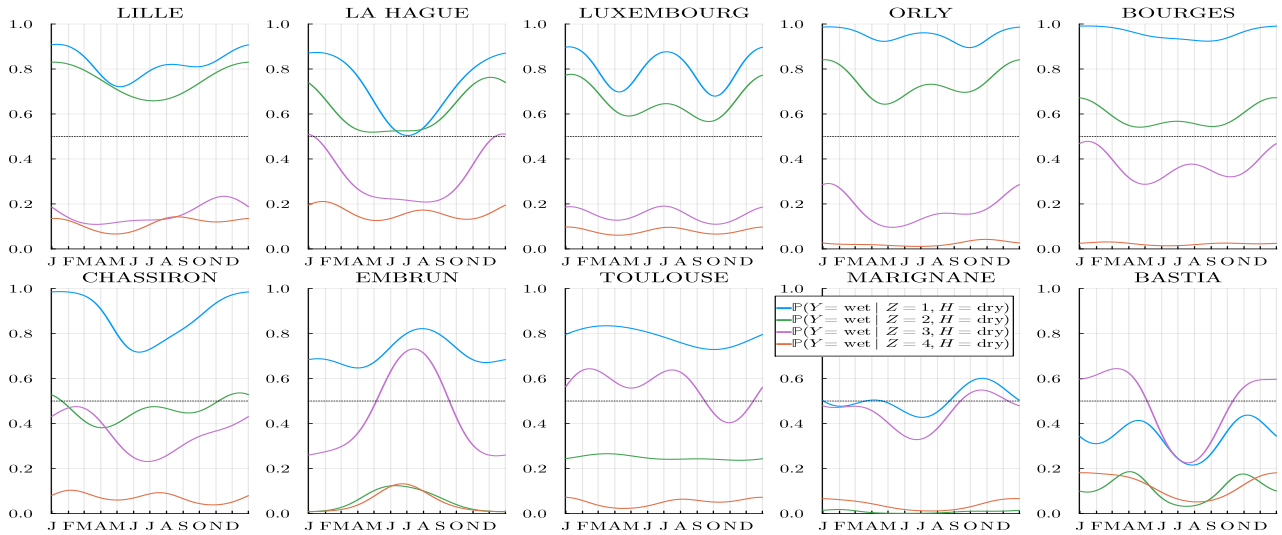


Figure 8: Estimated $\lambda_{k,t,s,h}$ probability for $m = 1$ and $h =$ dry, i.e., the probability of rain at the location s, conditionally to the hidden state $k \in [\![1, K]\!]$ and to a previous dry day. The stations are sorted by latitude from the northernmost (top left) to the southernmost (bottom right)

In almost all stations, the extreme states $Z = 1$ (4) are where it rains most (less) often. As we noticed in Figure 5, states 2 and 3 are different in the north and south.

4.3 Mean Rain Amount

Even if the model training does not involve any rain amounts $R_s^{(n)}$, the hidden states $Z = k$ should still be meaningful for these. In Figure 9, we plot the daily mean rain amount $R_{k,s}^{(t)} > 0$ for each station and climate state k. The values obtained are smoothed with periodic moving averaged of time window ± 15 days, see Appendix C for the definition. The "rainy" weather regime $k = 1$ is at almost every location and all year long, the state where it rains the most. Similarly, the "dry" regime $k = K$ is

where it rains the least. Interestingly, the intermediate regimes, $k = 2$, $k = 3$, are rainier in the north at different seasons. Southerner stations have a different behavior as expected.

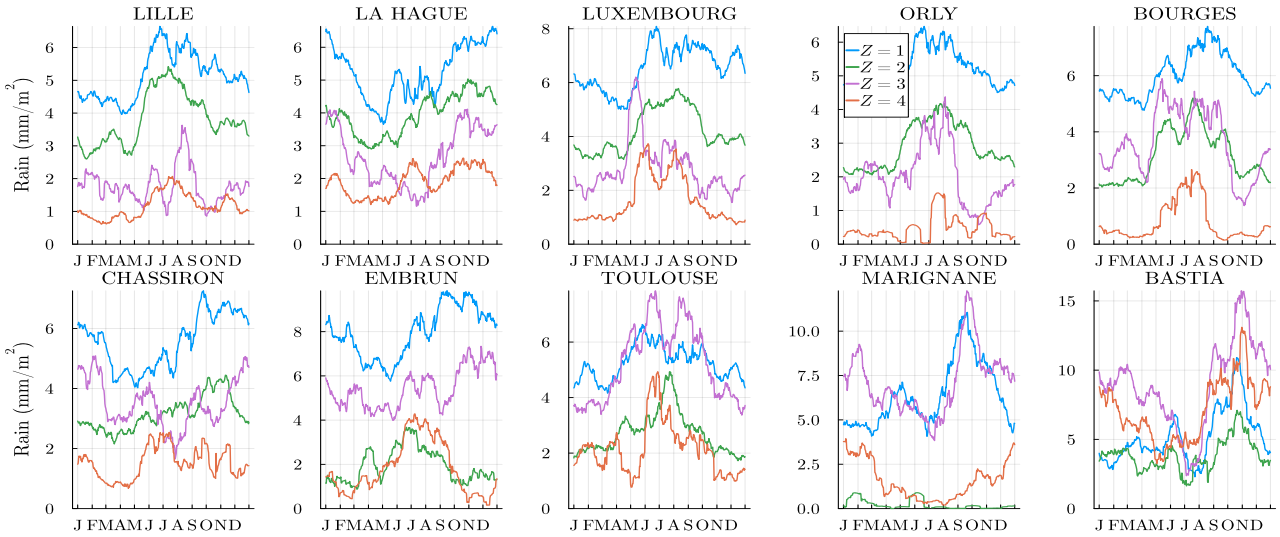


Figure 9: Daily mean strictly positive rain quantity $R > 0$ (mm) at every station per k^{th}-component. We smooth the results as in (11). We use the model $\mathcal{C}_1$, $K = 4$ to get a posteriori the most likely state associated with each date n, see (15).

4.4 Weather regime spells

To illustrate the dynamics of the weather regimes, we show in Figure 10 for different years the Viterbi estimated hidden states ($\hat{z}$) (see Section 3.3). As previously noticed, dry and wet spells last longer in

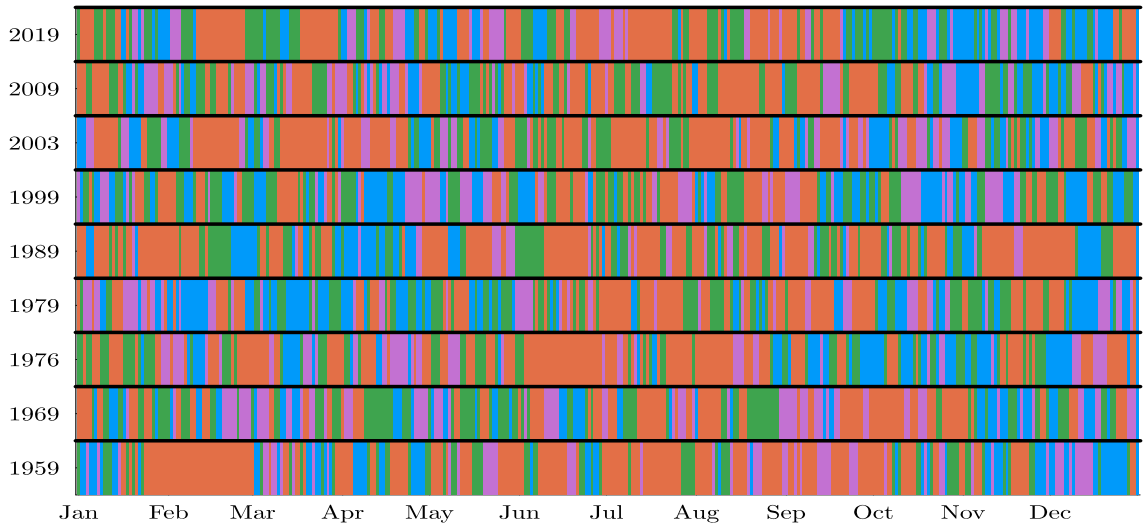


Figure 10: Estimated hidden states sequence for a selection of years. Each color corresponds to a hidden state, $Z = 1$ is blue, $Z = 2$ is green, $Z = 3$ is purple and $Z = 4$ is orange.

general than in other states. For historical events such as the drought of summer 1976, we observe a long dry sequence (27 days in a row in state $Z = 4$ starting from June 3rd). The famous 2003 heat wave from August 1st to August 15th also corresponds to a 15 day's dry spell.

5 Simulations: Multisite Rain Occurrence

Now that the model is fully inferred and interpreted, we will test its validity. To do so, we will sample multiple i.i.d. realizations of the training period 1956 to 2019 and compare several spatio-temporal statistics with the historical data.

5.1 Simulations Algorithm of the SHHMM

We first sample the hidden states $(z^{(n)} : n \in \mathcal{D})$ according to the nonhomogeneous periodic transition matrix Q_{t_n} and initial distribution ξ, then we draw the MRO $(y^{(n)} : n \in \mathcal{D})$ from the conditional emission distributions $f_{z^{(n)},t_n,s,h_s^{(n)}}$. The procedure is summarized, in Algorithm 1.

Algorithm 1: Simulation of the SHHMM

Result: Sequence hidden states $z^{(n)}$, sequence of MRO $y^{(n)}$
$z^{(n=1)} \sim \xi$;
for $n \in \mathcal{D}$ **do**
  $z^{(n)} \sim Q_{t_n}(z^{(n-1)}, \cdot)$;
end
$y^{(n=m-1:0)} = y_{\text{ini}}^{n=m-1:0}$;
for $n \in \mathcal{D}$ **do**
  for $s \in \mathcal{S}$ **do**
    $y_s^{(n)} \sim f_{z^{(n)},t_n,s,h_s^{(n)}}(\cdot)$;
  end
end

In the simulations, we choose the initial date as January 1, 1956. Our final date is Dec 31, 2019 so that the total simulated range is 64 years which corresponds to our data set span. We choose $\xi = (1, 0, 0, 0)$, i.e., $z^{(1)} = 1$ i.e., a rainy weather regime because it was a rainy day all over France on that day. We assume that the MRO prior to the first simulation day $y_{\text{ini}}^{n=m-1:0}$ are observed and use them as input to draw $y_s^{(1)} \sim f_{z^{(1)},t,s,h_s^{(1)}}$.

5.2 Results

In the following, we will use $M = 10^3$ i.i.d. realizations of the SHHMM over a 64-year span and compare its statistics with the 64 -year observed sequence.

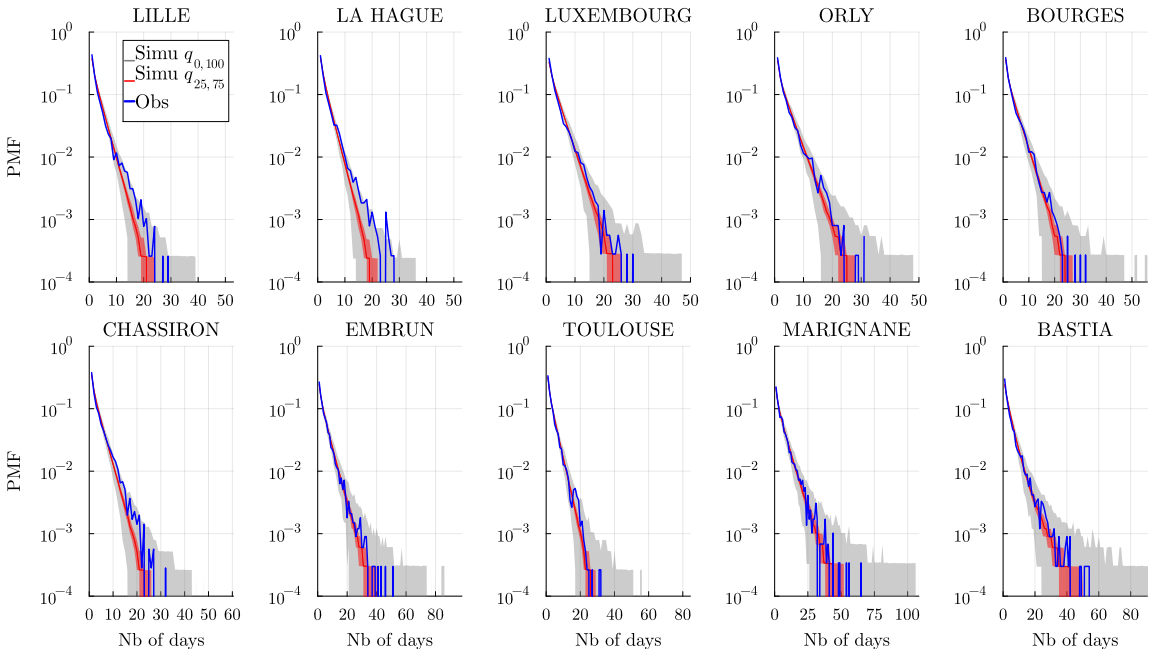


Figure 11: Dry spell distribution (in number of days) at every station and for a time range $\mathcal{D}$ of the historical data (blue) and of the $M = 10^3$ simulated wet spell distribution. The gray envelope covers the full range ($q_{0,100}$) of the simulations, while the red envelope covers the interquartile range ($q_{25,75}$) and the line is the median. Simulations are obtained over the same time range $\mathcal{D}$ and using the model $K = 4$, $d = 2$ and $\mathcal{C}_{m=1}$.

5.2.1 Dry/Wet state sequence

The dry spell sequences are of particular interest to estimate risk associated with droughts. We show the observed dry (wet) spells in Figure 11 (and 12) at all the stations and compare it to the simulated spells for the M realization. When the historical distribution is contained in the simulations' envelope, we may conclude that the model does a good job to reproduce the dry (wet) spells: note that this works systematically well, except for La Hague station at a few data points. It borders the Channel sea and is the northernmost station. Hence, it is not completely surprising that $m = 1$ local memory might not be enough to reproduce correctly its spells. At this station, a higher m might be required.

For the sake of comparison, in Appendix B, we show and discuss the distribution obtained using the memoryless $\mathcal{C}_{m=0}$ model to highlight the gain of the model $\mathcal{C}_{m=1}$ in both the center and the tails of spell distributions, see Figures 22 and 23. We note that even though wet spells are in general much shorter than dry spells, having $m = 1$ is necessary to reproduce accurately the wet spells.

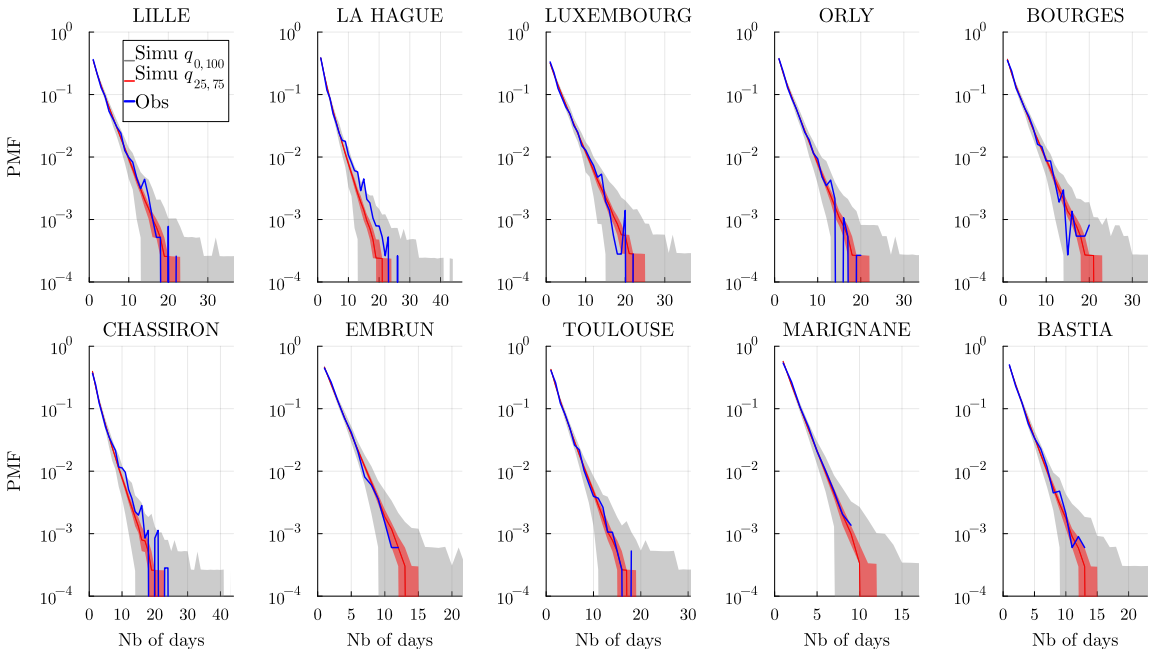


Figure 12: Wet spell distribution (in number of days) at every station and for a time range $\mathcal{D}$ of the historical data (blue) and of the $M = 10^3$ simulated wet spell distribution. The gray envelope covers the full range ($q_{0,100}$) of the simulations, while the red envelope covers the interquartile range ($q_{25,75}$) and the line is the median. Simulations are obtained over the same time range $\mathcal{D}$ and using the model $K = 4$, $d = 2$ and $\mathcal{C}_{m=1}$.

5.2.2 Spatial correlations

We compare in Figure 13 the observed and simulated $S(S-1)/2$ correlation coefficients between all sites $\text{cor}(\{Y_s^{(n)}\}_{n\in\mathcal{D}}, \{Y_{s'}^{(n)}\}_{n\in\mathcal{D}})$ for all $s \neq s' \in \mathcal{S}$. Most correlations are well reproduced, showing

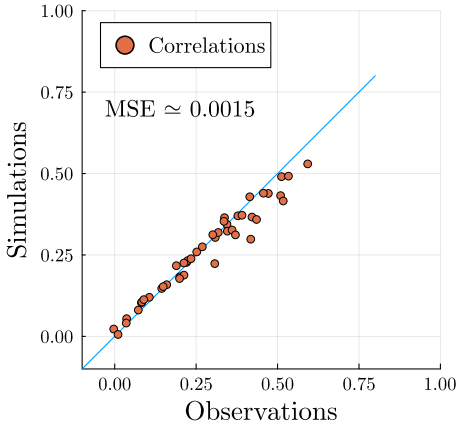


Figure 13: Observed pair correlations $\text{cor}(\{Y_s^{(n)}\}_{n\in\mathcal{D}}, \{Y_{s'}^{(n)}\}_{n\in\mathcal{D}})$ for all $s \neq s' \in \mathcal{S}$ compared with the correlations computed from the simulations (we average the $M = 10^3$ pair correlations of our simulations). The mean square error (MSE) with all correlation pairs is displayed on the figure.

that the conditional independence hypothesis 2.2 or 2.3 is empirically valid.

6 Modeling: Precipitations Amount

In this section, we attach to the model an add-on, a multisite precipitation amount generator. The procedure is done without modifying or re-training the original model. In fact, other variables such as Temperature, Solar Irradiance etc. could be attached similarly to what will be presented in this section. To do so, one only needs a generator for new variable e.g., AR(1) model for temperature, and allow its parameters to depend on the weather regimes $Z_n = k$ and to evolve smoothly (as in Section 2.4) with the day of the year t. Our hypothesis is that the new variable has some dependence on both the weather regime and the season. We discussed in Section 4, various spatiotemporal interpretations of the weather variable, thus it makes sense to consider how this global variable is relevant for other weather variables. Hence, the resulting add-on generator should generate a variable at least partially correlated with the original SHHMM. This makes the model very modular, allowing easy extensions without affecting its original performances and interpretations. Figure 9 highlights the rain amount dependence to the weather regime k and seasonality. This principle is applied in this section to build an add-on rainfall generator.

The Multisite Rain Amount (MRA in short) is denoted as

$$R^{(n)} := (R_1^{(n)}, \dots, R_S^{(n)}) \in \mathbb{R}_+^S.$$

Building directly an MRA generator is hard because of the ambivalent probabilistic nature of rain, being neither a discrete nor a continuous variable. Here we can just focus on strictly positive rain amounts $R > 0$ because the SHHMM directly tells when $R = 0$ or $R > 0$.

To train the rain amount generator, we will use the hidden states $Z^{(n)} = \hat{z}^{(n)}$ found in Section 3.3. The schematic of the resulting model is shown in Figure 14.

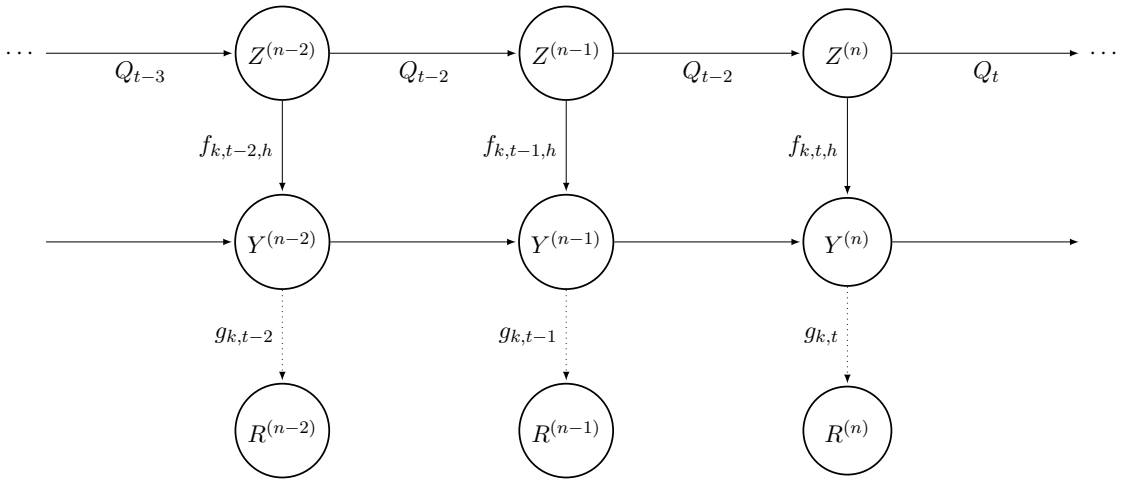


Figure 14: SHHMM Model with rain amounts. $g_{k,t}$ denotes the MRA generator with respect to the weather regime k and day of the year t.

6.1 Marginal rain distributions

The rain amount generator we use to fit the marginal distributions $R_s > 0$ at each station is a mixture $g(r)$ of two exponential distributions, with density

$$g(r) = w\frac{e^{-\frac{r}{\vartheta_1}}}{\vartheta_1} + (1-w)\frac{e^{-\frac{r}{\vartheta_2}}}{\vartheta_2}. \tag{9}$$

This choice is widely adopted in the literature e.g., [Kirshner, 2005, Touron, 2019b, Kroiz et al., 2020] and has only three parameters denoted by $\gamma = \{\vartheta_1, \vartheta_2, w\}$. It is flexible enough to be used for different climate types and locations. Other popular choices such as Gamma [Kroiz et al., 2020, Holsclaw et al., 2016] or heavy tails distributions [Baxevani and Lennartsson, 2015, Tencaliec et al., 2020] could be used at specific locations s or weather regimes k when needed. See [Chen and Brissette, 2014] for a review of univariate precipitation models. For example, precipitations in the south of France are less frequent than in the north but more intense, leading to extreme events which are better described with heavy-tailed distributions. In the simulation part Section 7, we show that despite being light-tailed, this choice of generator $g(r)$ trained w.r.t. weather regimes and seasonality is able to reproduce well both the bulk and the tails of most observed rain distribution.

As in Section 2.4, the parameters of the mixture are periodic functions $\gamma(t) = \{\vartheta_1(t), \vartheta_2(t), w(t)\}$, where $\vartheta_{1\,\text{or}\,2}(t) = e^{P_{1\,\text{or}\,2}(t)} > 0$, $w(t) = 1/(1+e^{P_\theta(t)})$ where the P functions are trigonometric polynomials (see Eq. (5)).

To fit the mixtures $g_{k,t,s}$ for each station s and hidden state k we use the classical **EM** algorithm. The maximization step has to be performed with numerical optimization as in Section 3. Note that, optimization can be done separately for each weather regime k and station s.

6.2 Multisite Distribution: Gaussian Copula

After training, the marginal distributions at each hidden state, and site $g_{k,t,s}$, we now focus on generating correlated Multisite Rainfall Amounts (MRA). To generate multisite rain occurrences (MRO), we used the conditional independence with respect to the hidden state (and possibly local history). For a vector of Bernoulli (dry/wet) random variables, this was enough to well approximate the observed correlation matrix (see Figure 13). However, for a vector of non-discrete random variables, such as rain amounts, mixtures of conditionally independent distributions typically underestimate the joint distribution [Holsclaw et al., 2016]. It means that despite the hidden states carrying some part of the MRA correlations, we have to add correlation through another way. A classical approach is to use copula [Nelsen, 2006]. Amongst the various families of copula, the Gaussian copula is the easiest to train and manipulate and has been used for weather models [Pandey et al., 2018, Kroiz et al., 2020]. In this paper, we will thus train and use Gaussian copula conditionally to the hidden states to generate multisite (strictly) positive rain amounts.

Let $(\rho_{s,s'})_{s,s'}$ be the correlations between a pair of stations s, s' for joint rainy events, i.e., $(\rho_{s,s'})_{s,s'} = \text{Cor}(R_s | R_s > 0, R_{s'} | R_{s'} > 0)$. To reproduce the correct observed (Pearson) correlation $(\rho)_{s,s'}$, we train a Gaussian copula. A Gaussian Copula takes a correlation matrix $\Sigma^{(G)} = \{\rho^{(G)}_{s,s'}\}_{s,s' \in \mathcal{S}^2}$ and the marginal distributions g_s as an input. The matrix $\Sigma^{(G)}$ is not directly observed, but for elliptic copula, there

is a relationship between the correlation $\rho^{(G)}$ and the Kendall (rank) correlations [Fang et al., 2002, Theorem 3.1],

$$\rho^{(G)} = \sin\left(\frac{\pi}{2}\rho_{\text{Kendall}}\right).$$

Hence, to compute $\rho^{(G)}$, we use the observed Kendall correlation ρ_{Kendall} which is preserved under monotonic transformation, such as quantile and CDF functions.

We estimate the correlation matrices $\Sigma_k^{(G)} = \{\rho_{k,s,s'}^{(G)}\}_{s,s'\in\mathcal{S}^2}$ conditionally to the hidden state $Z = k$. Indeed, we expect and observe that the weather regime impacts the correlation. For the driest state $Z = K$ rain event should be largely independent, in the rainy state precipitation should be correlated. We actually enforce the conditional independence when $k = K$ i.e., diagonal covariance matrix. This choice is also motivated by the lack of observations of joint rain events in state $k = K$.

In this work, we also assume for simplicity, that the correlation matrices have no seasonality dependence, i.e., independent on the day t. Moreover, we also do not model local temporal correlations for rain amounts. This shortcoming could be overcome using for example spatiotemporal covariance matrix [Benoit et al., 2018].

Simulation procedure. To simulate the rainfall amounts, we first simulate the SHHMM chain $(z^{(n)}, y^{(n)} : n \in \mathcal{D})$, see Algorithm 1. Then for all the stations where rain is predicted, $\mathcal{S}_{\text{wet}}^{(n)} = \{s : Y_s^{(n)} = \text{wet}, \forall s \in \mathcal{S}\}$, the rain amounts $R_s^{(n)} > 0$ are generated conditionally using the Gaussian copula with marginal $g_{z^{(n)},t_n,s}$ and correlation matrix $\Sigma^{(n)} = \{\rho_{z^{(n)},s,s'}\}_{s,s'\in(\mathcal{S}_{\text{wet}}^{(n)})^2}$.

Remark. In Appendix A, we show visually and with an approximate χ^2 test that the Gaussian copula model is a valid model for most station pairs. Note that this Gaussian copula can underestimate joint extreme rain amount [Renard and Lang, 2007] e.g., for close stations. In that case other copula might be used but will not be explored in this paper.

7 Simulations: Multisite Rain Amount

In this section, we will test the full multisite model combining the SHHMM and the rain amounts. We will test how marginal distributions, influence of seasonality over quantiles and correlations across stations are recovered by the model. Note that all previous results of the MRO simulations, see Section 5, are still valid since the addition of rain amount is done "on top" of the SHHMM.

In the simulations, we use the parameters obtained in Section 3.4: $m = 1$ local memory, $K = 4$ hidden states and $d = 2$ order of trigonometric polynomial. We use the SHHMM transition matrix and Bernoulli emission distributions obtained in Section 3.2 and the rain amount marginal and copula obtained in Section 6.

7.1 Correlations

We first compare the spatial correlation of MRA over the 64 years of data, i.e., for all pairs of stations s and s' we estimate $\text{Cor}(R_s, R_{s'})$. The results are shown in Figure 15 (left), where observed correlations

are compared to simulations. In Figure 15 (right), we perform a similar comparison for the symmetric tail correlation (or upper tail dependence) [Nelsen, 2006] defined by

$$(\rho_T)_{s,s'}(q) = \left((\rho_T)_{s|s'}(q) + (\rho_T)_{s'|s}(q)\right)/2 \tag{10}$$
$$\text{with } (\rho_T)_{s|s'}(q) = \mathbb{P}\left(R > F_{R_s}^{-1}(q) \mid R_{s'} > F_{R_{s'}}^{-1}(q)\right),$$

for $q \in [0,1]$. The tail correlation indicates how extreme events are correlated at different stations. We observe a good match for most stations, however for stations with larger tail correlation $\gtrsim 0.2$ the tail correlation is underestimated by the simulations. This can be an indication that the Gaussian copula is not enough for these pairs of stations. Improvement using Student copula (which manipulation is less easy but more capable to generate tail dependence) is a possibility that should be explored in future work.

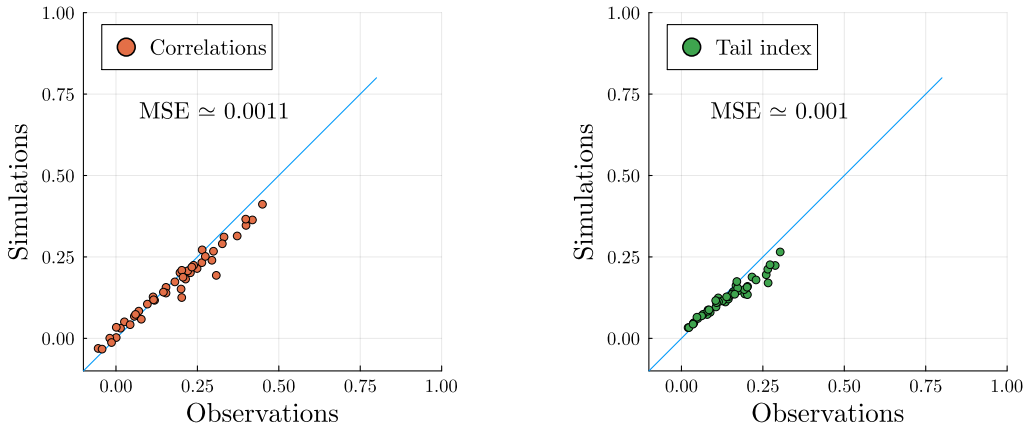


Figure 15: Comparison of the multisite correlation (left) and symmetric tail correlation (10) with $q = 0.95$ (right) from the observed and simulated data. The correlations computed from simulations are averaged over $M = 10^3$ realizations.

7.2 Rain Amounts

Distribution of precipitations. We show the nonzero rain amount distributions $R_s > 0$ at each station s (accumulated across the 64 years of data) in Figure 16. It shows the historical distributions (blue) and 10^3 realizations of our generator (gray).

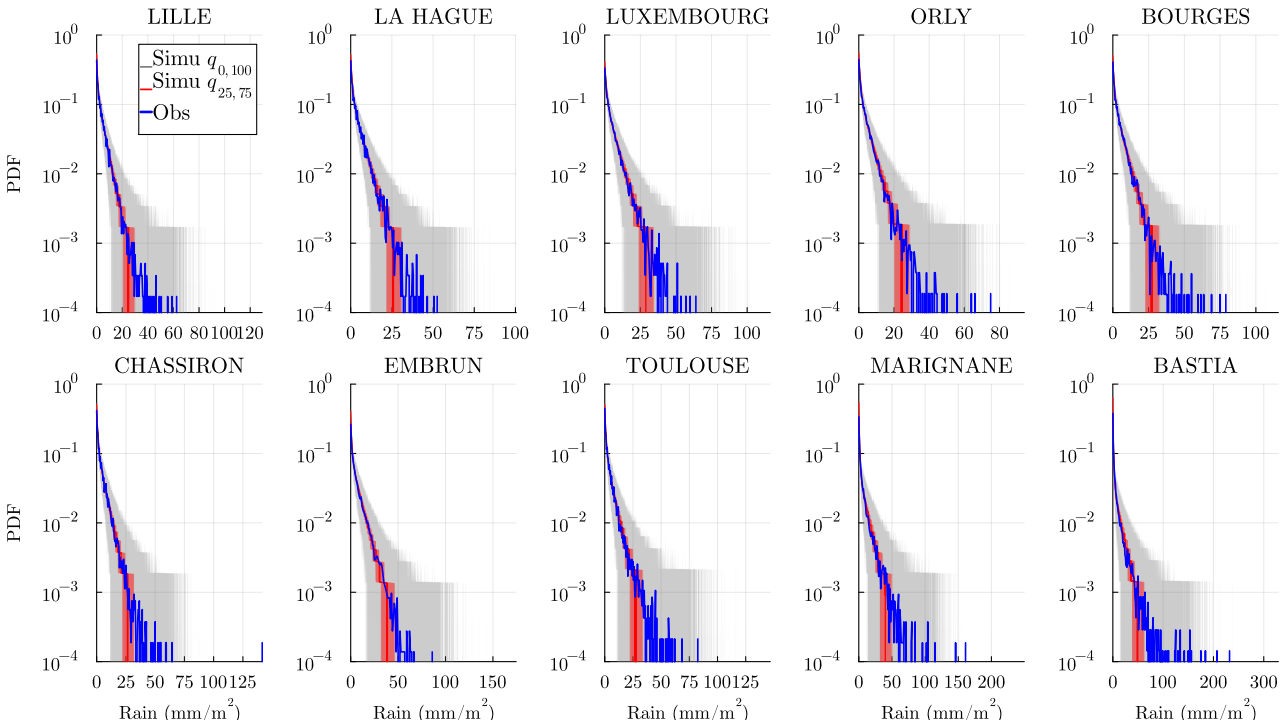


Figure 16: Distribution of the nonzero precipitation amount $R > 0$ (mm/m^2) at every station and for all years $\mathcal{T}$. Historical data (blue line). The gray envelope covers the full range ($q_{0,100}$) of the $M = 10^3$ simulations, while the red envelope covers the interquartile range ($q_{25,75}$) and the line is the median.

The model reproduces the bulk of the distributions as well as the tails. Even for some stations in the south (Toulouse, Marignane, Bastia) where the precipitation PDF has heavier tails, our model is able to capture extremes. This might be due to the seasonal training of the marginals Eq. (9) allowing the distributions to be more extreme in late summer when heavy storms are common. However, at some stations like Luxembourg, it can generate extremes twice as large as the current maximal value observed, which might be questionable.

Precipitations during the year. To test the seasonality of the model, we show the quantiles $0.1, 0.5, 0.9$ of the accumulated monthly amount at every location.

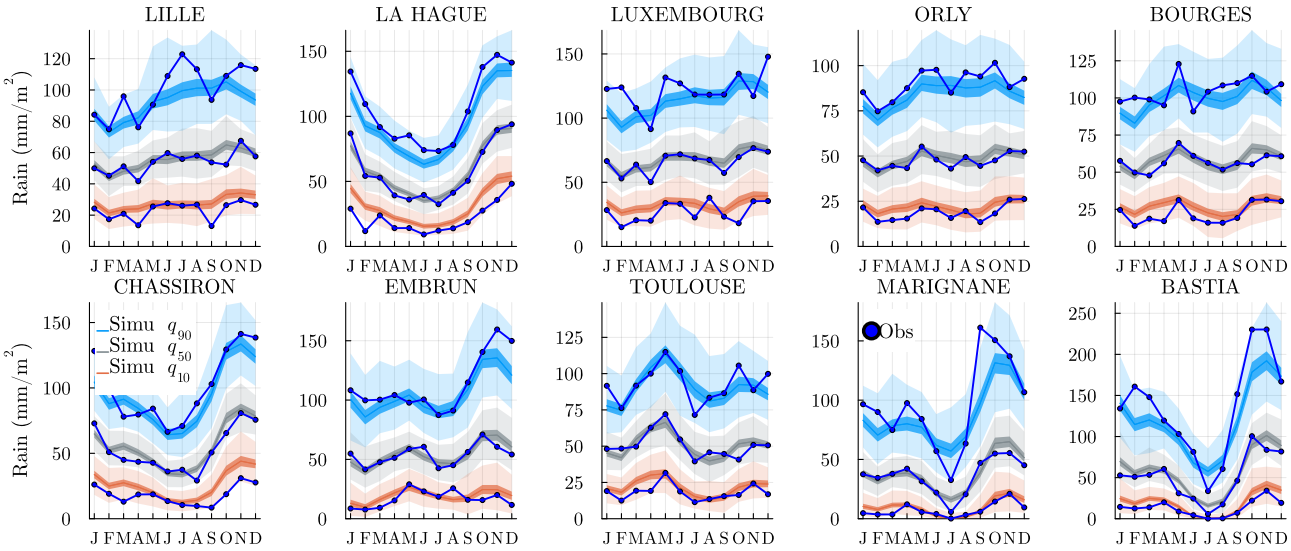


Figure 17: $(0.1, 0.5, 0.9)$–quantile of the cumulated monthly mean rain amount mm/month in (orange, gray, light blue) respectively. Historical data (dark blue). For each quantile we show the envelope of the $M = 10^3$ simulations, and in darker colors the $[25, 75]$ percentiles and the median.

This type of figure checks how our model performs regionally, over a 30-day period and in different regimes (very dry month, median and rainy month). Most observed points are located in the envelope of our $M = 10^3$ simulations.

8 Application to climate change projections

So far, we have trained and validated the SHHMM model using historical data. Using the same hyperparameters as found in the model selection, see Section 3.4, we can train the model on other data sets. In this section, we show how the Stochastic Weather Generator developed in this paper can be used to study climate change impacts. The focus of the paper is not to perform an in depth analysis, but rather to show as a *proof of concept* how SWG could be useful in that context. To this end, we will train the model with projection data made available by climate model institutes participating in the scientific projects coordinated in the IPCC framework [Arias et al., 2021]. A new Coupled Model Intercomparison Project is launched for each new IPCC cycle, and each participating institute run the newest versions of their global climate model or earth system model under prescribed radiative forcing conditions. Because these simulations are global and present biases compared to local observations, we will use the downscaled and bias adjusted projections provided by the French climate service DRIAS. It is based on a selection of regional projections made in the international CORDEX initiative based on CMIP5 global projections (projections made in the framework of the 5th IPCC assessment report).

8.1 DRIAS data

We use the DRIAS website [Soubeyroux et al., 2021] that aggregates different regional (European) projections forced by some chosen global projections made by different institutes. DRIAS-2020 provides thirty climate projections (2006-2100) with 3 scenarios (RCP2.6, RCP4.5 and RCP8.5) and twelve historical simulations (1951-2005). These simulations are further downscaled and bias adjusted over France using the SAFRAN reanalysis [Vidal et al., 2010] covering France with 8 km resolution for many daily variables. We select the closest grid points to the $S = 10$ considered stations and extract the precipitation amount. The exact grid point choice should not matter too much, since the reanalyzed simulations are smoothly interpolated. Since these physical models tend to overestimate the frequency of light rain amounts R, we set to $R = 0$ all amounts smaller than 0.1 mm to match what is done at the experimental weather stations.

8.2 Direct comparison of model on the reference period

Climate models provide historical simulations (1951-2005) to be able to validate models against observed data. Because a model does not simulate the same interannual variability as observed, the evaluations are based on the statistical properties of the variables rather than on their chronology.

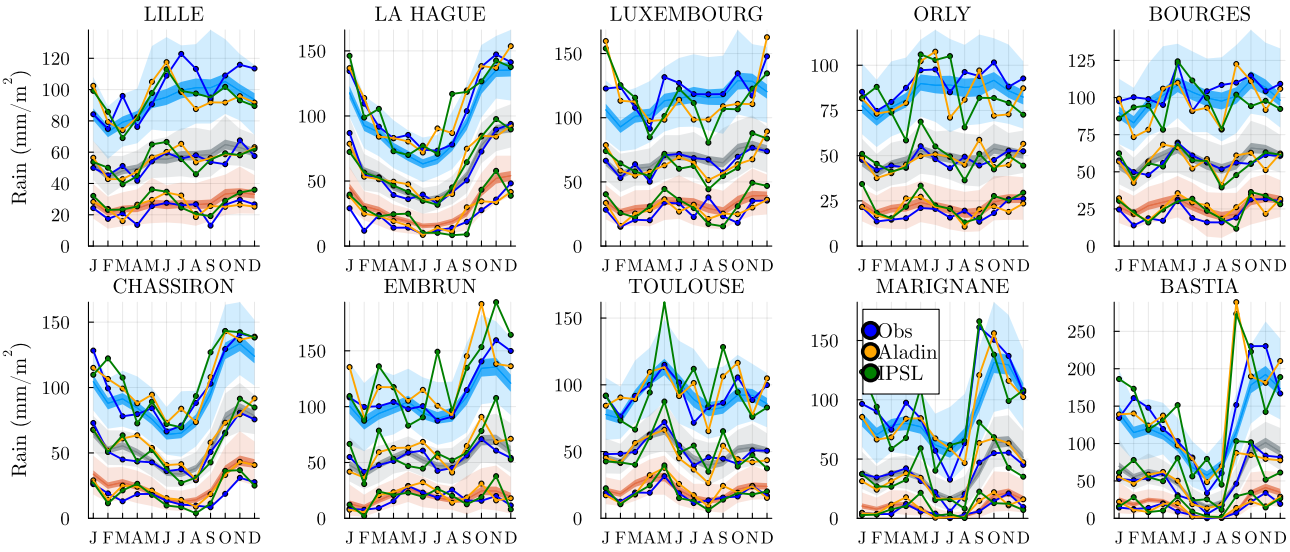


Figure 18: Same as Figure 17 with added monthly rain quantile for the model Aladin (CNRM-ALADIN63 - CNRM-CERFACS-CNRM-CM5) and IPCC (IPSL-WRF381P - IPSL-IPSL-CM5A-MR) on the reference period 1952 to 2006.

For example, in Figure 18, we compare the monthly rain quantiles computed as in Section 7.2 obtained from the historical climate simulations and from the SHHMM simulations (the same as previously trained on historical observation). We use here two climate models Aladin (CNRM-ALADIN63 - CNRM-CERFACS-CNRM-CM5) and IPSL (IPSL-WRF381P - IPSL-IPSL-CM5A-MR) as an example. Using a SWG allows a better sampling of the natural climate variability because it is possible to run much more realizations than can be done with climate models. This sample can then be used to check

how climate model simulations are positioned. For example at Lille station, in July for the 0.9 quantile we observe that the historical point at $\simeq 120\,\mathrm{mm/m^2}$ (blue) is far from the two climate models (orange and green) at $\simeq 100\,\mathrm{mm/m^2}$. However, when looking at the predicted statistical envelope, the climate models are exactly at the median while the historical observation is actually an extreme value. When comparing the two climate models, we observe that the IPSL model produces more points outside the statistical envelope than the Aladin model, suggesting that the model may present stronger biases.

8.3 Training on RCP scenarios

Once the comparison has been made for the historical period, in this section, we will study how the spatial rainfall may evolve in the future, by fitting the SHHMM on climate model projections under different RCP scenarios. The RCP scenarios are designed to represent differentiated trajectories of greenhouse gas and aerosol emissions that drive climate change until the end of the century (and beyond in some cases). To do so, we select the data over a 64 years range, here 2032-2096, which simplifies the statistical comparison with the 64 years range of the historical data we considered.

In Figure 19 the transition matrix obtained when training on historical and IPSL-RCP8.5 data are compared. The aim here is only to highlight the ability of the SHHMM to be used in climate change conditions, not to conduct an impact study, that's why only one climate model is used. The two matrices are still close, which tells that the hidden states of our model are robust to parameter evolutions. However, we can observe interesting differences. For example, $Q_{3\to 3}$ and $Q_{1\to 1}$ are significantly larger in summer months. Weather regime 1 and 3 were interpreted as rainy all over France and heavy rain in the south respectively. This means that the IPSL model under RCP8.5 projects longer stretches of heavy rain. Figure 20, shows the analog of Figure 17 with simulations from the trained SHHMM model with IPSL-RCP8.5 data and in blue the historical data. It clearly shows that the IPSL model under RCP8.5 scenario projects rainier periods. In fact, it is known that the regional climate model EURO-Coordinated Regional Downscaling Experiment used in all DRIAS models present this type of bias [Boé et al., 2020, Vautard et al., 2021]. In particular, summer periods are consistently rainier, even for the 0.1-quantile of the monthly mean rain amount.

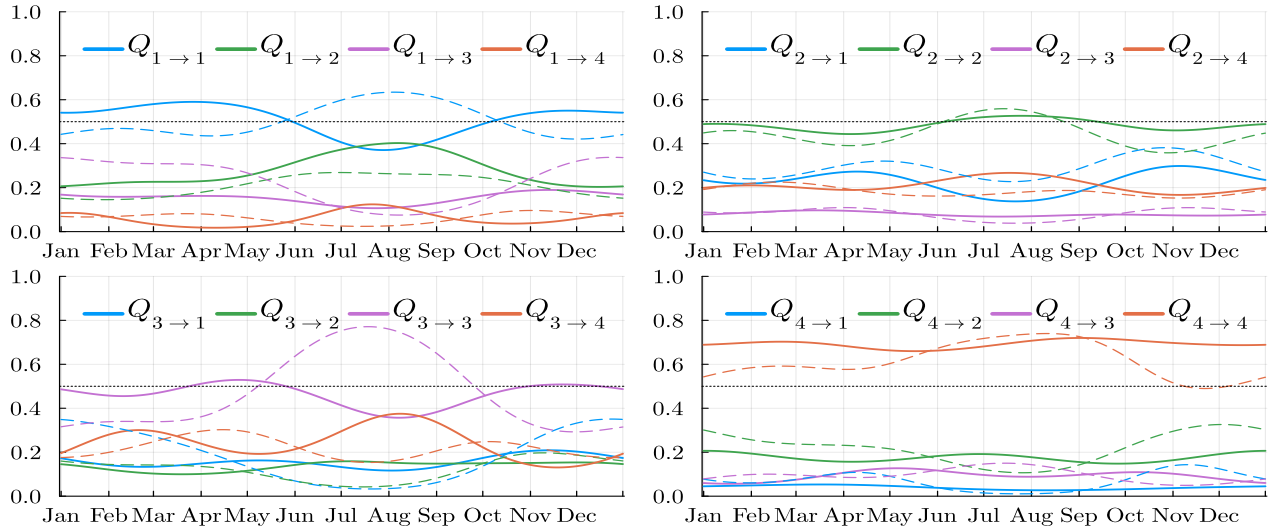


Figure 19: Temporal variation of the transition matrix Q_t trained on historical data (plain line) and on RCP8p5 from IPSL-WRF381P data (dotted line) for the period 2032-2096.

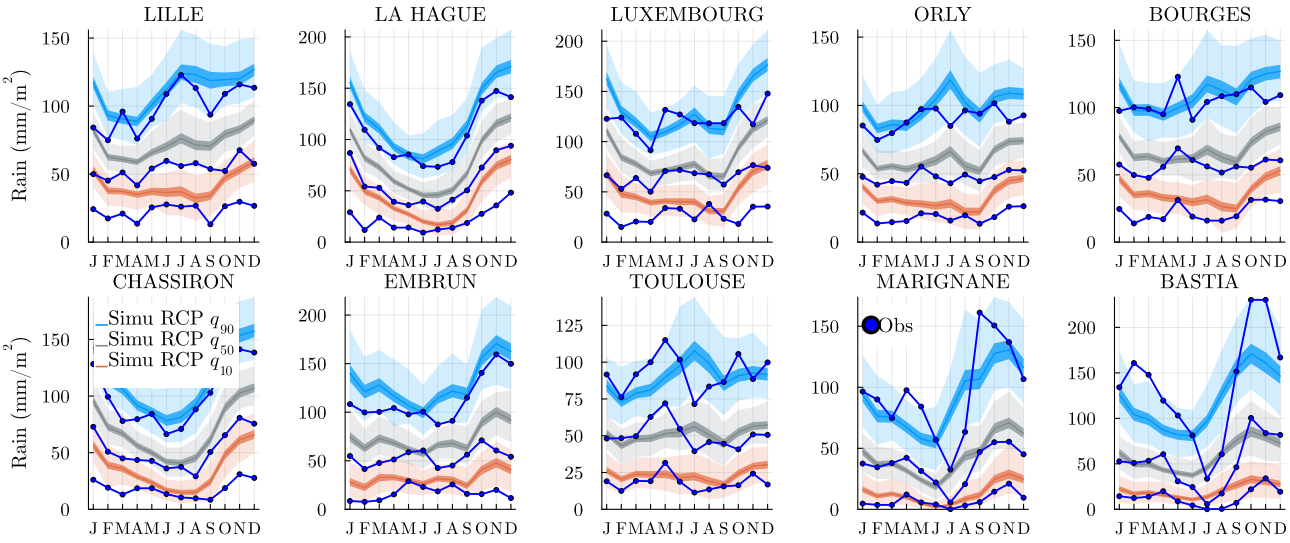


Figure 20: $(0.1, 0.5, 0.9)$–quantile of the cumulated monthly mean rain amount mm/month in (orange, gray, light blue) respectively. Historical data (dark blue) and $M = 10^3$ realization of the model trained under with the RCP8p5-IPSL-WRF381P data for the period 2032-2096. For each quantile we show the envelope of the M simulations, and in darker colors the $[25, 75]$ percentiles and the median.

This example shows how the proposed SWG can be used to analyze and compare models: either directly interpreting the coefficients changes, or sampling from the fitted model to study extreme behaviors.

9 Conclusion

In this paper, we define a multisite Stochastic Weather Generator for precipitation named Seasonal Hierarchical Hidden Markov Model (SHHMM). Even though it is based on a Hidden Markov Model introduced for weather applications in the 90s, we propose an original training method based on two important assumptions (a) conditional independence for the Multivariate Rain Occurrence (MRO) variable (see Eqs. (2)-(4)), (b) Imposed smooth seasonal evolution of most model parameters (see Section 2.4). Assumption (a) forces the model to learn spatial correlations leading to fully interpretable hidden states (weather regimes). This is different from what is usually done, where hidden states and correlation coefficients are trained together for continuous variables. Thanks to the discrete nature of MRO and the station locations, we checked the validity of hypothesis (a), see Figure 13. The other assumption (b) is natural and has been introduced before in [Touron, 2019a], it stabilizes training i.e., removes a lot of identifiability issues that occur while training non-homogenous HMM and leverages for the relatively small number of observation year. To facilitate the training, we also introduced in Section 3.1 a naive estimate for the SHHMM that is used as the initial state of the Baum Welch Expectation Maximization algorithm (see Section 3.2). To capture more of the local weather i.e., station wise, we introduced a hierarchical dependence of Rain Occurrence with their past weather, allowing better temporal correlations. The model selection was performed using the Integrated Complete-data Likelihood criteria, leading in particular to the selection of four hidden states interpreted extensively as France wise Weather Regimes in Section 4. In particular, we were able to showcase how the hidden states found are similar to the four Euro-Atlantic weather regimes commonly defined in meteorology. The model was extensively tested with simulations. Its performances in terms of reproductions of dry/wet spells and precipitation amount is very good even in the distributions tails. Moreover, the model structure allows very easily to add other weather variable on top of the HMM without modifying the hidden states. In fact, new variables like rain amounts benefits from the trained hidden states and are adjusted with respect to them. Eventually, we showed how this generator can be used with climate change projections, either to interpret model parameters changes or resample from these simulations.

Many small improvements could be considered, such as different models for rain amount or local memory at different locations to account for regional specificities. In fact, even the station locations and number could be optimized to satisfy better hypothesis (a). A temporal correlation with previous rain amount is also possible with spatiotemporal correlation matrix [Benoit et al., 2018]. Extending the model with new weather variables such as temperature on top of the current model (and its hidden states) is the next major challenge to be considered. Indeed, if the Weather Regimes found here are surely relevant for temperature, evapotranspiration, solar radiations etc., they are also probably not enough to fully correlate all variables. Moreover, the problematic of downscaling [Vrac et al., 2007, Holsclaw et al., 2016] i.e., having finer resolution (denser station distributions) around an area of interest should be tackled with a similar spirit i.e., new stations are fitted and correlated on top of the current model. Finally, application to study climate change projections e.g., comparison and exploration of extremes should be explored in depth and is left for further work.

Appendix

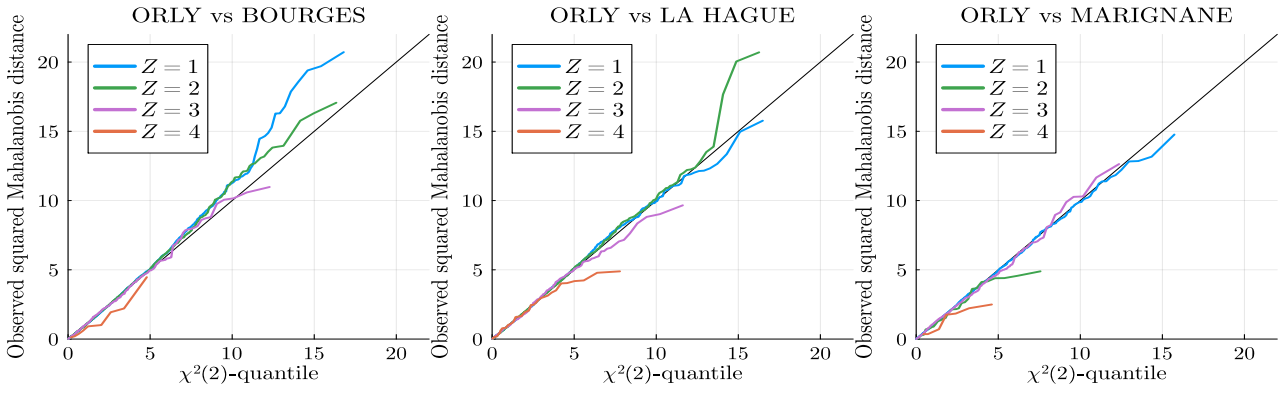
A Gaussian Copula

To check the Gaussian copula approximation for the joint rain events between station pairs, we transform our data into an empirical bivariate distribution with Normal margin to test its quantiles against the one of a true bivariate normal distribution. Note that these checks are mostly qualitative since we apply the procedure to time series, meaning we are outside the i.d.d. framework where these kinds of tests are valid.

In detail, given a pair of stations (s, s') and a hidden state $Z = k$, we consider the joint positive rain amount $R_{s,s',k} = (R_s > 0, R_{s'} > 0) \mid Z = k$ for all dates n, so we can remove the superscript n. We first have to transform the observations to pseudo observation, i.e. $R_{s,s',k} \in \mathbb{R}_+^2 \xrightarrow{\eta} (u_{s,k}, u_{s',k}) \in [0,1]^2$. There are several possible transformations η e.g., the estimated marginal CDF or ordinal ranking. We use the latter one as done in the package `Copulas.jl` [Laverny and Jimenez, 2024], that we use for all our Copulas simulations. This pseudo observations are then transformed to Normal distributions using the transformation $X_{s,s',k} = (\phi^{-1}(u_{s,k}), \phi^{-1}(u_{s',k}))$ where ϕ^{-1} is the quantile function of the standard Normal distribution. For a vector $x \in \mathbb{R}^n$ and a $n \times n$-correlation matrix Σ_M, the squared Mahalanobis distance is defined, as

$$D_M(x) = x^\intercal \Sigma_M^{-1} x.$$

We use the correlation coefficients $\rho^{(G)}_{s,s',k}$ obtained in Section 6.2 to build the 2×2-matrix Σ_M and compute the Mahalanobis distance for all samples. For a true bivariate normal distribution, the distribution of D_M follows a $\chi^2(\nu)$-distribution with $\nu = 2$ degree of freedom. In Figure 21, we compare the quantile of the $\chi^2(\nu = 2)$-distribution with the $D_M(X_{s,s',k})$ for two pairs of stations and each hidden state $Z = k$. Note that we simplify the analysis, considering only one covariance matrix instead of the four fitted in Section 6.2.



(a) 182 km (b) 334 km (c) 587 km

Figure 21: Three examples of qq-plot to test the Gaussian copula hypothesis. Is shown the Mahalanobis distances between stations pairs vs. $\chi^2(\nu = 2)$ distribution. A good match means that the Gaussian copula hypothesis to generate pairs $(R_s > 0, R_{s'} > 0) \mid Z = k$ is well satisfied.

The correspondence is good even for close pairs. It means that Gaussian copulas are adapted when

stations are far enough apart. For $Z = 1$, i.e., the rainiest weather, only 3 out of 45 station pairs fail the one-sided Kolmogorov-Smirnov test with 95% confidence level that compares the theoretical $\chi^2(\nu = 2)$ distribution with the observed squared Mahalanobis distance. These are the pairs Bourges-Orly, Lille-Orly and Lille-Luxembourg which are amongst the closest pairs, e.g., see Figure 21a. Interestingly, for a slightly bigger distance, the pair (334 km) Orly-La Hague passes the test (see Figure 21b). This clearly indicates anisotropy in the correlation repartition. For other weather regimes $Z > 1$, the Gaussian copula hypothesis also works well in most stations with enough data.

B Comparison with memoryless model $\mathcal{C}_{m=0}$

In Section 2.2 and 2.3 we defined respectively model $\mathcal{C}_{m=0}$ and $\mathcal{C}_{m>0}$. We later selected $\mathcal{C}_{m=1}$ using the ICL criteria, see Figure. 4. We show here the performances of the $\mathcal{C}_{m=0}$ model in terms of dry/wet spell on Figures 22 and 23. The observed distribution is shown, while the $M = 10^3$ simulation quantile envelope are displayed. These figures are to be compared with the Figures 11 and 12 produced by the $\mathcal{C}_{m=1}$ model. In the bulk of the spell distributions i.e., short spells with higher probability, the difference is important (note that the log scale tends to minimize visually the effect), e.g., Embrun and Marignane for the dry spells and all wet spells distributions. This indicates that the model $\mathcal{C}_{m=0}$ without local memory overestimate very short wet spells (and dry spell in a lesser measure). At some stations it also underestimates the longer spells (tails) e.g., Bastia for wet spells.

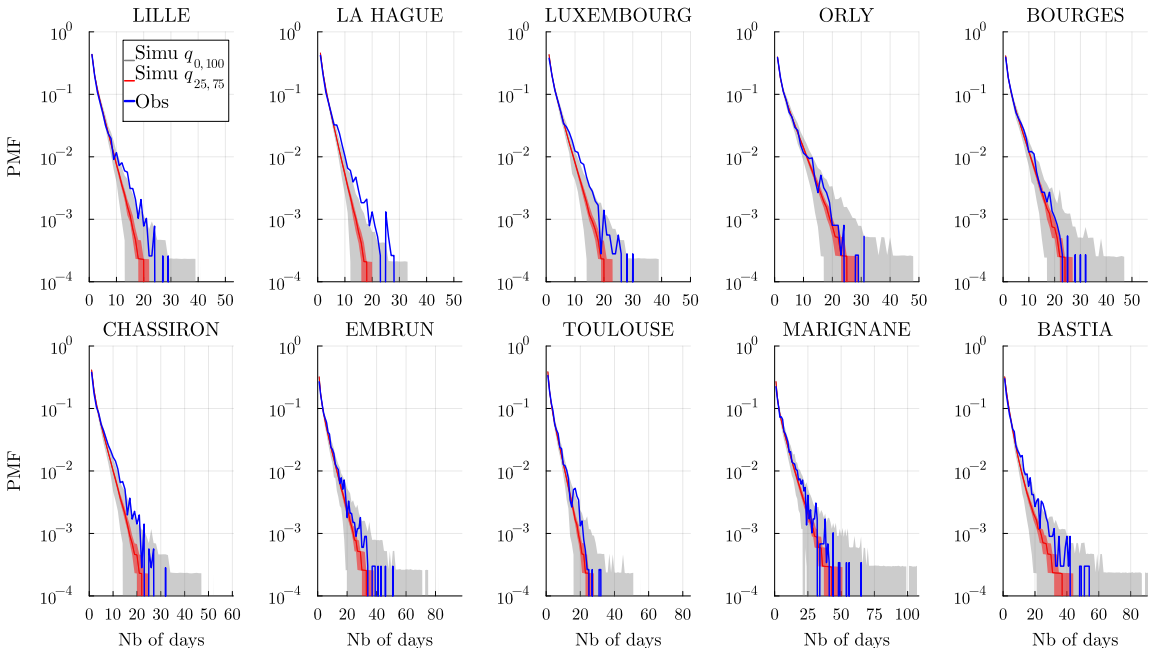


Figure 22: Dry spell distribution (in number of days) at every station and for a time range $\mathcal{D}$ of the historical data (blue) and the $M = 10^3$ simulated wet spell distribution. The gray envelope covers the full range ($q_{0,100}$) of the simulations, while the red envelope covers the interquartile range ($q_{25,75}$) and the line is the median. Simulations are obtained over the same time range $\mathcal{D}$ and using the memory less model $K = 4$, $d = 2$ and $\mathcal{C}_{m=0}$.

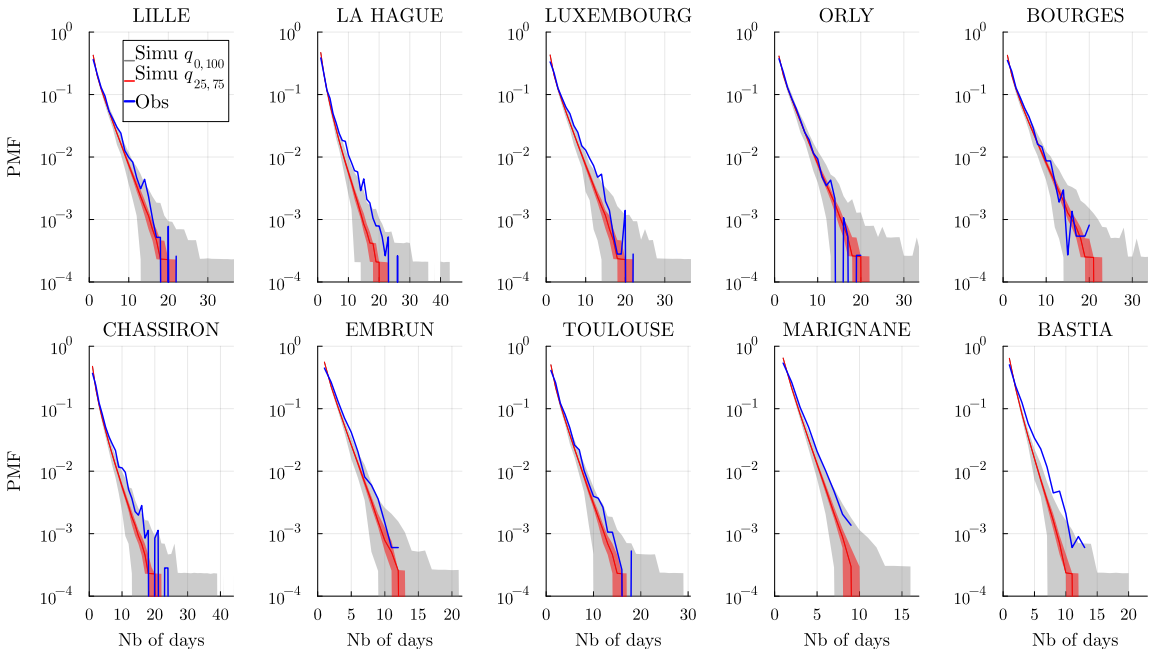


Figure 23: Wet spell distribution (in number of days) at every station and for a time range $\mathcal{D}$ of the historical data (blue) and the $M = 10^3$ simulated wet spell distribution. The gray envelope covers the full range ($q_{0,100}$) of the simulations, while the red envelope covers the interquartile range ($q_{25,75}$) and the line is the median. Simulations are obtained over the same time range $\mathcal{D}$ and using the memory less model $K = 4$, $d = 2$ and $\mathcal{C}_{m=0}$.

C Periodic moving average

We define the periodic moving average used in Figure 9. Given a T-periodic observable $X^{(t)}$ for $t \in \mathcal{T}$, the associated moving average $\bar{X}^{(t)}$ is given by

$$\bar{X}^{(t)} = \sum_{h=-H}^{h=H} \mathcal{K}\left(\frac{h}{H}\right) \frac{X^{(t+h)}}{\sum_{h=-H}^{h=H} \mathcal{K}\left(\frac{h}{H}\right)}, \tag{11}$$

where $t \in [\![1, T]\!]$, $X^{(t \pm T)} = X^{(t)}$ and $\mathcal{K}$ is a kernel. In the paper, we choose the window size $H = 15$ with the kernel is the Epanenchikov kernel $\mathcal{K}(u) = \frac{3}{4}(1-u)^2 \mathbf{1}_{|u| \leq 1}$ and $T = 366$.

D Baum-Welch algorithm for Seasonal Hierarchical HMM

We use the same model $\mathcal{C}_{m>0}$ as described in Section 2 and will describe the inference procedure using the Baum-Welch algorithm for Seasonal Hierarchical HMM (SHHMM).

We recall that θ stands for all the SHHMM $(\xi, Q_t, f_{k,t,h})_{k \in \mathcal{K}, t \in \mathcal{T}, h \in \mathcal{H}^{(m)}}$ model parameters (see

Section 2.3). To fit the model, we must find the θ maximizing the observed likelihood,

$$\begin{aligned}
L_\theta\left(y^{(1:N)}\right) &= \mathbb{P}_\theta\left(Y^{(1:N)} = y^{(1:N)}\right) \\
&= \sum_{z_1,\dots,z_n} \mathbb{P}_\theta\left(Y^{(1:N)} = y^{(1:N)}, Z^{(1:N)} = z^{(1:N)}\right) = \sum_{z_1,\dots,z_n} L_\theta\left(y^{(1:N)}, z^{(1:N)}\right) \\
&= \sum_{z_1,\dots,z_n} f_{z_N,t_N}\left(y^{(N)} \mid h^{(N)}\right) \mathbb{P}_\theta\left(Y^{(1:N-1)} = y^{(1:N-1)}, Z^{(1:N)} = z^{(1:N)}\right) \\
&= \sum_{z_1,\dots,z_n} f_{z_N,t_N}\left(y^{(N)} \mid h^{(N)}\right) Q_{z_{N-1},z_N}(t_N) \mathbb{P}\left(Y_{1:N-1}, Z_{1:N-1} = z_{1:N-1}\right) \\
&= \sum_{z_1,\dots,z_N} \xi_{z_1,h_1} f_{z_1,t_1}\left(y^{(1)} \mid h^{(1)}\right) \prod_{n=2}^{N} Q_{z_{n-1},z_n}(t_n) f_{z_n,t_n}\left(y^{(n)} \mid h^{(n)}\right),
\end{aligned}$$

where the index $z_n \in [\![1, K]\!]$ for all $n \in \mathcal{D}$.

The Baum-Welch algorithm is an iterative Expectation Maximization, where the likelihood is increased sequentially, i.e., at a step (i) of the algorithm $L_{\theta^{(i)}} \leq L_{\theta^{(i+1)}}$.

Let us detail the procedure and show that the classical element of the Baum Welch algorithm for a (homogeneous) HMM proof remains valid when considering a SHHMM. The first step is to consider the conditional expectation of the loglikelihood of the parameter θ, L_θ, with respect to the parameter at step (i), $\theta^{(i)}$, i.e.,

$$\begin{aligned}
R\left(\theta, \theta^{(i)}\right) &= \mathbb{E}_{\theta^{(i)}}\left[\log L\left(Y^{(1:N)}, Z^{(1:N)}; \theta\right) \mid Y^{(1:N)}\right] \\
&= \sum_{k,l=1}^{K} \sum_{n=1}^{N-1} \pi^{\theta^{(i)}}_{n,n+1|n}(k,l) \log Q_{t_n}(k,l) + \sum_{k=1}^{K} \sum_{n=1}^{N} \pi^{\theta^{(i)}}_{n|N}(k) \log f_{k,t_n}(y^{(n)} \mid h^{(n)}) \\
&\quad + \sum_{k=1}^{K} \pi^{\theta^{(i)}}_{1|n}(k) \log \xi_k.
\end{aligned}$$

where the $\pi^{\theta^{(i)}}_{n|N}(k)$ and $\pi^{\theta^{(i)}}_{n,n+1|N}(k,l)$ are the smoothing probabilities computed with the current $\theta^{(i)}$ and defined as follows

$$\pi_{n|N}(k) = \mathbb{P}_\theta\left(Z^{(n)} = k \mid Y^{(1:N)}\right), \quad \forall n \in [1, N],$$

$$\pi_{n,n+1|N} = \mathbb{P}_\theta\left(Z^{(n)} = k, Z^{(n+1)} = l \mid Y^{(1:N)}\right), \quad \forall n \in [1, N-1].$$

These probabilities can be computed using the Forward-Backward procedure which is also valid for Periodic Hierarchical HMM.

The **E** and **M** steps alternate as follows:

(1) **Initialization.** We initialize the algorithm with an initial HMM of parameter $\theta^{(0)}$.

(2) **E**-step: Compute $R(\theta, \theta^{(i)})$, it corresponds here to get the smoothing probabilities for the current parameter $\theta^{(i)}$.

(3) **M**-step: Maximize $R(\theta, \theta^{(i)})$ with respect to θ. Due to the sum expression of R, this step can be done independently for each parameter $\theta = (\xi, Q, f)$. In particular, one can update the emissions distributions $f_{t,k}(y_n \mid h_n)$ independently of the transition matrix. If we don't assume a periodic parametric form for the transition matrices, we can maximize explicitly each Q_t independently.

(4) Step **E** and **M** are repeated alternatively until the observed likelihood has converged to a local maximum.

Fundamental inequality of the EM algorithm. To prove that increasing $R(\theta \mid \theta^{(i)})$ also increases the observed likelihood $L_\theta\left(Y^{(1:N)}\right)$ we first rewrite the observed likelihood as

$$\log L_\theta\left(Y^{(1:N)}\right) = \log L_\theta\left(Y^{(1:N)}, Z^{(1:N)}\right) - \log L_\theta\left(Z^{(1:N)} \mid Y^{(1:N)}\right).$$

The conditional expectation of L_θ with respect to the current parameter $\theta^{(i)}$, for all $\theta, \theta^{(i)}$ gives

$$\begin{aligned}
\mathbb{E}_{\theta^{(i)}}\left[\log L_\theta\left(Y^{(1:N)}\right) \mid Y^{(1:N)}\right] &= \log L_\theta\left(Y^{(1:N)}\right) \\
&= \mathbb{E}_{\theta^{(i)}}\left[\log L_\theta\left(Y^{(1:N)}, Z^{(1:N)}\right) - \log L_\theta\left(Z^{(1:N)} \mid Y^{(1:N)}\right) \mid Y^{(1:N)}\right] \\
&= R\left(\theta, \theta^{(i)}\right) - \sum_{Z^{(1:N)}} \mathbb{P}_{\theta^{(i)}}\left(Z^{(1:N)} \mid Y^{(1:N)}\right) \log \mathbb{P}_\theta\left(Z^{(1:N)} \mid Y^{(1:N)}\right) \\
&= R\left(\theta, \theta^{(i)}\right) + \mathcal{R}\left(\theta, \theta^{(i)}\right).
\end{aligned}$$

The Gibbs's inequality ensures that $\mathcal{R}\left(\theta, \theta^{(i)}\right) \geq \mathcal{R}\left(\theta^{(i)}, \theta^{(i)}\right)$, so that we obtain

$$\log L_\theta\left(Y^{(1:N)}\right) - \log L_{\theta^{(i)}}\left(Y^{(1:N)}\right) \geq R\left(\theta, \theta^{(i)}\right) - R\left(\theta^{(i)}, \theta^{(i)}\right).$$

Hence, when we maximize (or increase) $R\left(\theta, \theta^{(i)}\right)$ with respect to θ we also increase the observed loglikelihood.

Smoothing and filtering probabilities. The smoothing probabilities can be expressed as

$$\begin{aligned}
\pi_{n|N}(k) &= \mathbb{P}_\theta\left(Z^{(n)} = k \mid Y^{(1:N)} = y^{(1:N)}\right) = \frac{\mathbb{P}_\theta\left(Z^{(n)} = k, Y^{(1:N)} = y^{(1:N)}\right)}{\mathbb{P}_\theta\left(Y^{(1:N)} = y^{(1:N)}\right)} \\
&= \frac{\mathbb{P}_\theta\left(Z^{(n)} = k, Y^{(1:n)} = y^{(1:n)}\right) \mathbb{P}_\theta\left(Y^{(n+1:N)} = y^{(n+1:N)} \mid Z^{(n)} = k, Y^{(1:n)} = y^{(1:N)}\right)}{\mathbb{P}_\theta\left(Y^{(1:N)} = y^{(1:n)}\right)} \\
&= \frac{\alpha_n(k)\beta_n(k)}{\sum_{l=1}^K \alpha_n(l)\beta_n(l)},
\end{aligned}$$

with

$$\begin{aligned}
\alpha_n(k) &= \mathbb{P}_\theta\left(Z^{(n)} = k, Y^{(1:n)} = y^{(1:n)}\right), \\
\beta_n(k) &= \mathbb{P}_\theta\left(Y_{n+1:N} = y^{(n+1:N)} \mid Z^{(n)} = k, Y^{(1:n)} = y^{(1:n)}\right) \\
&= \mathbb{P}_\theta\left(Y^{(n+1:N)} = y^{(n+1:N)} \mid Z^{(n)} = k, Y^{(n-m+1:n)} = y^{(n-m+1:n)}\right).
\end{aligned}$$

Similarly,

$$\pi_{n,n+1|N}(k,l) = \mathbb{P}_\theta\left(Z^{(n)}=k, Z^{(n+1)}=l \mid Y^{(1:N)}=y^{(1:N)}\right) = \frac{\alpha_n(k)\beta_{n+1}(l) f_{t_{n+1},l}(y^{(n+1)} \mid h^{(n+1)}) Q_t(k,l)}{\mathbb{P}_\theta\left(Y^{(1:N)}=y^{(1:N)}\right)}.$$

Forward-Backward procedure. The forward α, backward β variables are computed iteratively

$$\begin{cases} \alpha_1(k) = f_{k,t_1}(y^{(1)} \mid h^{(1)})\xi_k, \\ \alpha_n(k) = f_k(y^{(n)} \mid h^{(n)}) \sum_{l=1}^K Q_{t-1}(l,k)\alpha_{n-1}(l), \quad \text{for } 1 < n \leq N, \end{cases}$$

$$\begin{cases} \beta_N(k) = 1, \\ \beta_n(k) = \sum_{l=1}^K f_{t_{n+1},l}(y^{(n+1)} \mid h^{(n+1)}) Q_t(k,l)\beta_{n+1}(l), \quad \text{for } 1 \geq n < N. \end{cases}$$

E Initialization of the HMM fitting: The slice estimate algorithm

In Section 3.2, we use the slice estimate to initialize the Baum Welch algorithm with parameters $\theta^{(0)}$. We detail in this Appendix the inference of this slice estimate. Indeed, a random choice of $\theta^{(0)}$ into the Baum Welch algorithm could lead to bad local maxima or longer convergence time.

E.1 The EM algorithm for the emission distributions

The 64 years of data provides on each day t a sample of size 64, considered independent and identically distributed. Hence, for each $t \in \mathcal{T}$, independently of each other, we use a standard EM algorithm to fit the emissions distribution $\{f_{1,t}, \cdots, f_{K,t}\}$. For a given date t e.g., February 28, the samples will consist of all Feb 28 from the data set, i.e., from year 1956 to year 2019. The ensemble of date n corresponding to the same day t is denoted $\mathcal{N}_t$. To enrich each of these small datasets, we add the observations of every $t \pm 6, 12$ day to each $\mathcal{N}_t$ (with periodicity $T = 366$). These additional days should come from very similar distributions as the one from date t, as assumed by the smoothness assumption, see Section 2.4, but also be far enough to be considered as independent samples. For our current dataset, each day t has for samples all the dates n with associated $t_n \in \{t, t \pm 6, t \pm 12\}$ which gives $|\mathcal{N}_t^+| = 320$ samples for each[2], where we denote by $\mathcal{N}_t^+$ the enriched dataset.

On a day t, the mixture probability writes for an observation vector, $y = (y_1, \cdots, y_S)$ with history $h = (h_1, \cdots, h_S)$ as

$$f_t(y \mid h) = \mathbb{P}\left(Y^{(n)} = y \mid H^{(n)} = h\right) = \sum_{k=1}^K \mathbb{P}\left(Z^{(n)}=k\right)\mathbb{P}\left(Y^{(n)}=y \mid H^{(n)}=h, Z^{(n)}=k\right)$$
$$= \sum_{k=1}^K \pi_{k,t} \prod_{s=1}^S \mathbb{P}\left(Y_s^{(n)} = y_s \mid H_s^{(n)} = h_s, Z^{(n)}=k\right) = \sum_{k=1}^K \pi_{k,t} \prod_{s=1}^S f_{k,t,s}(y_s \mid h_s)$$

[2]Except for February 29 (and Feb 17, 23 and March 6,12)

where we use the conditional independence, see Section 2.3, and denoted the weight $\mathbb{P}\left(Z^{(n)}=k\right)$ by $\pi_{k,t}$. The parameters to fit are the mixture weights $\pi_{k,t}$ and the Bernoulli parameters $\lambda_{k,t,h,s}$ for $k \in \mathcal{K}$, $s \in \mathcal{S}$ and $h \in \mathcal{I}_s^c$. We denote with a hat the estimated parameters $\widehat{\pi}_{k,t}$ and $\widetilde{\theta}_{k,t,h,s}$.

E.2 Algorithm

The different steps, Expectation (**E**) and Maximization (**M**) of the algorithm are standard. The mixture to fit is composed of products of Bernoulli distribution conditionally on the history vector h. The same mixtures appear a lot in classification problems, for example for digits reconnaissance [Bishop, 2006, Section 9.3.3.].

E.3 Random initialization

We choose randomly 10 random initial parameters $(\pi_{k,t}^{(0)}, \lambda_{k,t,h,s}^{(0)})$, run the algorithm and select the converged point with the largest observed likelihood, defined as

$$\begin{aligned}\ell_{\text{slice}}(y \mid h; \widetilde{\theta}_{k,t,s,h}, \widehat{\pi}_{k,t}) &= \log\left(\prod_{n \in \mathcal{N}_t^+} \mathbb{P}\left(Y^{(n)}=y^{(n)} \mid H^{(n)}=h^{(n)}\right)\right)^{N_t} \\ &= \sum_{n \in \mathcal{N}_t^+} \log\left(\sum_{k=1}^{K} \widehat{\pi}_{k,t} \prod_{s=1}^{S} \widetilde{f}_{k,t,s}(y_s^{(n)} \mid h_s^{(n)})\right),\end{aligned}$$

where $\widetilde{f}_{k,t,s}$ denotes the distribution with the estimated parameters $\widetilde{\theta}_{k,t,s,h}$.

E.4 Ordering the Hidden States

A mixture distribution is identifiable up to relabeling of its components, meaning the mixture defined by $(\pi_k, \lambda_{k,t,h,s})$ cannot be distinguished from the mixture $(\pi_{\sigma(k)}, \lambda_{\sigma(k),t,h,s})$ where σ is a permutation of $\mathcal{K}$. In our case, we need to ensure that the parameters evolve coherently with t so that labels k always refer to the same hidden states. To do so, we select one reference station in our study, Bourges, and for all $t \in \mathcal{T}$ relabel as follows:

- **Model $\mathcal{C}_0$:** Sort the probability of rain for $k \in \{1, \cdots, K\}$ from the lowest to the largest at the reference station Bourges

$$\widetilde{\theta}_{k=(1),t,s=\text{Bourges}} > \widetilde{\theta}_{k=(2),t,s=\text{Bourges}} > \cdots > \widetilde{\theta}_{k=(K),t,s=\text{Bourges}}.$$

- **Model $\mathcal{C}_m$:** Sort the probability of rain for $k \in \{1, \cdots, K\}$ conditionally to the driest history variable $h_d = (d, \cdots, d)$ from the lowest to the largest at the reference station Bourges

$$\widetilde{\theta}_{k=(1),t,s=\text{Bourges},h_d} > \widetilde{\theta}_{k=(2),t,s=\text{Bourges},h_d} > \cdots > \widetilde{\theta}_{k=(K),t,s=\text{Bourges},h_d}.$$

This sorting provides a natural interpretation to each hidden state: $k=(1)$ corresponds to a "rainy" climate where the probability of rain is the largest in Bourges and hopefully in the rest of the *métropole*

(continental France). The $k = (K)$ state corresponds to a "dry" climate where the probability of no rain is the largest.

The choice of Bourges to extract the hidden variable is heuristically justified by the fact that this station is located roughly at the center of the geographic area under study, and its parameters $\widetilde{\theta}_{k,t,s=\text{Bourges},h_d}$ are well separated for different k.

E.5 Transition matrices

To finish the SHHMM inference, we estimate the transition matrices $Q(t)$. To do so we will first infer the filtered probability of all hidden states given the model and observations using $\widetilde{f}_{k,t}$ and $\widetilde{\pi}_{k,t}$,

$$\gamma_k^{(n)} = \mathbb{P}\left(Z^{(n)} = k \mid Y^{(n)} = y^{(n)}, H^{(n)} = h^{(n)}\right) = \frac{\widetilde{\pi}_k \prod_{s=1}^{S} \widetilde{f}_{k,t,s}(y_s^{(n)} \mid h_s^{(n)})}{\sum_{l=1}^{K} \widetilde{\pi}_l \prod_{s=1}^{S} \widetilde{f}_{l,t,s}(y_s^{(n)} \mid h_s^{(n)})}. \tag{15}$$

The maximum *a posteriori* estimator is, then

$$\widetilde{z}^{(n)} = \text{argmax}_{k \in \{1,\cdots,K\}} \gamma_k^{(n)}.$$

This yields the sequence of hidden states $\{Z^{(n)} : n \in \mathcal{D}\}$. The transition matrices can be estimated by counting the number of transitions on a day t from a state k to l divided by the total number of transitions from k,

$$\widetilde{Q}_{kl}(t) = \frac{\sum_{n \in \mathcal{N}_t} \mathbf{1}_{\widetilde{z}^{(n)}=k, \widetilde{z}^{(n+1)}=l}}{\sum_{n \in \mathcal{N}_t} \sum_{l=1}^{K} \mathbf{1}_{\widetilde{z}^{(n)}=k, \widetilde{z}^{(n+1)}=l}}.$$

E.6 Multiple random initialization

To prevent the EM procedure to reach an irrelevant local minimum, we run 10 time the algorithm with added noise around the initial state. For each coefficient $c \in \theta^{(slice)}$, we randomize as $c^{\text{rand}} = c\,(1 + \sigma Z)$, where we take $\sigma = 0.5$ and $Z \sim \mathcal{N}(0, 1)$.

E.7 Slice Estimate Initialization VS. Naive Random Initialization

We show her the improvement given by the Slice estimate compare to pure random initialization. The loglikelihood obtained with this initialization is compared with 10 pure random initialization, where all the $\beta \sim \mathcal{N}(0, 0.1)$. The relative improvement is plotted in Figure 24. The slice estimate always gives greater or equal final loglikelihood for all tested models. It was strictly greater in 35 out of 42 models tested and equal in the remaining 7 cases, which include small models with $K = 2$ and 3 where inference is expected to be easier.

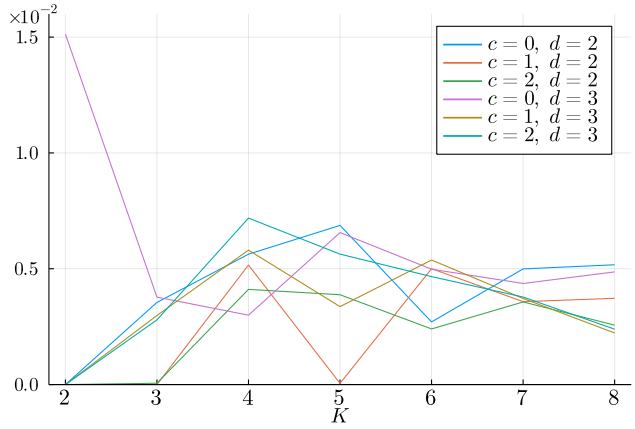


Figure 24: Relative improvement $(L_{\text{rand}} - L_{\text{slice}})/|L_{\text{slice}}|$ of the final loglikelihood obtained with random or slice initialization. The slice estimate always gives better or equal loglikelihood.

Note that even an improvement of less than a percent can lead to quite different models, in particular regarding interpretability of the hidden states.

References

[Ailliot et al., 2015a] Ailliot, P., Allard, D., Monbet, V., and Naveau, P. (2015a). Stochastic weather generators: an overview of weather type models. Journal de la société française de statistique, 156(1):101–113.

[Ailliot et al., 2015b] Ailliot, P., Bessac, J., Monbet, V., and Pène, F. (2015b). Non-homogeneous hidden Markov-switching models for wind time series. Journal of Statistical Planning and Inference, 160:75–88.

[Ailliot et al., 2020] Ailliot, P., Boutigny, M., Koutroulis, E., Malisovas, A., and Monbet, V. (2020). Stochastic weather generator for the design and reliability evaluation of desalination systems with Renewable Energy Sources. Renewable Energy, 158:541–553.

[Ailliot and Monbet, 2012] Ailliot, P. and Monbet, V. (2012). Markov-switching autoregressive models for wind time series. Environmental Modelling & Software, 30:92–101.

[Ailliot et al., 2009] Ailliot, P., Thompson, C., and Thomson, P. (2009). Space–time modelling of precipitation by using a hidden Markov model and censored Gaussian distributions. Journal of the Royal Statistical Society: Series C (Applied Statistics), 58(3):405–426.

[Allman et al., 2009] Allman, E. S., Matias, C., and Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. The Annals of Statistics, 37(6A):3099–3132.

[Arias et al., 2021] Arias, P., Bellouin, N., Coppola, E., Jones, R., Krinner, G., Marotzke, J., Naik, V., Palmer, M., Plattner, G.-K., Rogelj, J., Rojas, M., Sillmann, J., Storelvmo, T., Thorne, P.,

Trewin, B., Achuta Rao, K., Adhikary, B., Allan, R., Armour, K., Bala, G., Barimalala, R., Berger, S., Canadell, J., Cassou, C., Cherchi, A., Collins, W., Collins, W., Connors, S., Corti, S., Cruz, F., Dentener, F., Dereczynski, C., Di Luca, A., Diongue Niang, A., Doblas-Reyes, F., Dosio, A., Douville, H., Engelbrecht, F., Eyring, V., Fischer, E., Forster, P., Fox-Kemper, B., Fuglestvedt, J., Fyfe, J., Gillett, N., Goldfarb, L., Gorodetskaya, I., Gutierrez, J., Hamdi, R., Hawkins, E., Hewitt, H., Hope, P., Islam, A., Jones, C., Kaufman, D., Kopp, R., Kosaka, Y., Kossin, J., Krakovska, S., Lee, J.-Y., Li, J., Mauritsen, T., Maycock, T., Meinshausen, M., Min, S.-K., Monteiro, P., Ngo-Duc, T., Otto, F., Pinto, I., Pirani, A., Raghavan, K., Ranasinghe, R., Ruane, A., Ruiz, L., Sallée, J.-B., Samset, B., Sathyendranath, S., Seneviratne, S., Sörensson, A., Szopa, S., Takayabu, I., Tréguier, A.-M., van den Hurk, B., Vautard, R., von Schuckmann, K., Zaehle, S., Zhang, X., and Zickfeld, K. (2021). Technical summary. In Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J., Maycock, T., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., editors, Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, pages 33–144. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

[Baxevani and Lennartsson, 2015] Baxevani, A. and Lennartsson, J. (2015). A spatiotemporal precipitation generator based on a censored latent Gaussian field. Water Resources Research, 51(6):4338–4358.

[Bellone et al., 2000] Bellone, E., Hughes, J. P., and Guttorp, P. (2000). A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts. Climate research, 15(1):1–12.

[Benoit et al., 2018] Benoit, L., Allard, D., and Mariethoz, G. (2018). Stochastic Rainfall Modeling at Sub-kilometer Scale. Water Resources Research, 54(6):4108–4130.

[Bishop, 2006] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Information science and statistics. Springer, New York.

[Boé et al., 2020] Boé, J., Somot, S., Corre, L., and Nabat, P. (2020). Large discrepancies in summer climate change over Europe as projected by global and regional climate models: Causes and consequences. Climate Dynamics, 54(5):2981–3002.

[Cappé et al., 2005] Cappé, O., Moulines, E., and Rydén, T. (2005). Inference in hidden Markov models. Springer Series in Statistics. Springer, New York.

[Cassou, 2004] Cassou, C. (2004). Du changement climatique aux régimes de temps : l'oscillation nord-atlantique [prix prud'homme 2002]. La Météorologie, 2004(45):21–32.

[Celeux and Durand, 2008] Celeux, G. and Durand, J.-B. (2008). Selecting hidden Markov model state number with cross-validated likelihood. Computational Statistics, 23(4):541–564.

[Chen and Brissette, 2014] Chen, J. and Brissette, F. P. (2014). Stochastic generation of daily precipitation amounts: Review and evaluation of different models. Climate Research, 59(3):189–206.

[Christophe and Pompili, 2018] Christophe, P. and Pompili, B. (2018). Rapport fait au nom de la commission d'enquête sur "la sûreté et la sécurité des installations nucléaires". Assemblée Nationale, 1122. http://www.assemblee-nationale.fr/15/rap-enq/r1122-tI.asp.

[Cowpertwait et al., 2007] Cowpertwait, P., Isham, V., and Onof, C. (2007). Point process models of rainfall: developments for fine-scale structure. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 463(2086):2569–2587.

[de Chaumaray et al., 2023] de Chaumaray, M. D. R., Kolei, S. E., Etienne, M.-P., and Marbac, M. (2023). Estimation of the Order of Non-Parametric Hidden Markov Models using the Singular Values of an Integral Operator.

[Diaconis and Freedman, 1980] Diaconis, P. and Freedman, D. (1980). Finite Exchangeable Sequences. The Annals of Probability, 8(4):745 – 764.

[ECAD, 2022] ECAD (2022). Home European Climate Assessment & Dataset. https://www.ecad.eu/.

[Fang et al., 2002] Fang, H.-B., Fang, K.-T., and Kotz, S. (2002). The Meta-elliptical Distributions with Given Marginals. Journal of Multivariate Analysis, 82(1):1–16.

[Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27.

[Gyllenberg et al., 1994] Gyllenberg, M., Koski, T., Reilink, E., and Verlaan, M. (1994). Non-uniqueness in probabilistic numerical identification of bacteria. Journal of Applied Probability, 31(2):542–548.

[Hersbach et al., 2020] Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N. (2020). The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, 146(730):1999–2049.

[Holsclaw et al., 2016] Holsclaw, T., Greene, A. M., Robertson, A. W., and Smyth, P. (2016). A bayesian hidden Markov model of daily precipitation over South and East Asia. Journal of Hydrometeorology, 17(1):3–25.

[Hughes and Guttorp, 1994a] Hughes, J. P. and Guttorp, P. (1994a). A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. Water resources research, 30(5):1535–1546.

[Hughes and Guttorp, 1994b] Hughes, J. P. and Guttorp, P. (1994b). Incorporating spatial dependence and atmospheric data in a model of precipitation. Journal of Applied Meteorology and Climatology, 33(12):1503–1515.

[Hughes et al., 1999] Hughes, J. P., Guttorp, P., and Charles, S. P. (1999). A non-homogeneous hidden Markov model for precipitation occurrence. Journal of the Royal Statistical Society: Series C (Applied Statistics), 48(1):15–30.

[Kirshner, 2005] Kirshner, S. (2005). Modeling of multivariate time series using hidden Markov models. PhD thesis, University of California, Irvine. PhD thesis in Information and Computer Science, http://www.datalab.uci.edu/papers/kirshner_thesis.pdf.

[Kroiz et al., 2020] Kroiz, G. C., Basalyga, J. N., Uchendu, U., Majumder, R., Barajas, C. A., Gobbert, M. K., Markert, K., Mehta, A., and Neerchal, N. K. (2020). Stochastic precipitation generation for the Potomac river basin using hidden Markov models. UMBC Physics Department, http://hpcf-files.umbc.edu/research/papers/CT2020Team1.pdf.

[Laverny and Jimenez, 2024] Laverny, O. and Jimenez, S. (2024). Copulas.jl: A fully Distributions.jl-compliant copula package. Journal of Open Source Software, 9(94):6189.

[McLachlan and Krishnan, 2007] McLachlan, G. J. and Krishnan, T. (2007). The EM Algorithm and Extensions. John Wiley & Sons.

[Métivier, 2024] Métivier, D. (2024). StochasticWeatherGenerators.jl.

[Miloshevich et al., 2024] Miloshevich, G., Lucente, D., Yiou, P., and Bouchet, F. (2024). Extreme heat wave sampling and prediction with analog Markov chain and comparisons with deep learning. Environmental Data Science, 3:e9.

[Monbet and Ailliot, 2017] Monbet, V. and Ailliot, P. (2017). Sparse vector Markov switching autoregressive models. Application to multivariate time series of temperature. Computational Statistics & Data Analysis, 108:40–51.

[Nelsen, 2006] Nelsen, R. B. (2006). An Introduction to Copulas. Springer Series in Statistics. Springer, New York, NY.

[Pandey et al., 2018] Pandey, P. K., Das, L., Jhajharia, D., and Pandey, V. (2018). Modelling of interdependence between rainfall and temperature using copula. Modeling Earth Systems and Environment, 4(2):867–879.

[Pascual et al., 2017] Pascual, D., Pla, E., Fons, J., and Abdul-Malak, D. (2017). Climate change impacts on water availability and human security in the intercontinental biosphere reserve of the mediterranean (Morocco-Spain). In Environmental Change and Human Security in Africa and the Middle East, pages 75–93. Springer.

[Pawlowsky-Glahn and Buccianti, 2011] Pawlowsky-Glahn, V. and Buccianti, A. (2011). Compositional data analysis: Theory and applications. John Wiley & Sons.

[Pohle et al., 2017] Pohle, J., Langrock, R., van Beest, F. M., and Schmidt, N. M. (2017). Selecting the Number of States in Hidden Markov Models: Pragmatic Solutions Illustrated Using Animal Movement. Journal of Agricultural, Biological and Environmental Statistics, 22(3):270–293.

[Ranger et al., 2022] Ranger, N. A., Mahul, O., and Monasterolo, I. (2022). Assessing financial risks from physical climate shocks: A framework for scenario generation. World Bank, Washington, DC. https://hdl.handle.net/10986/37041.

[Renard and Lang, 2007] Renard, B. and Lang, M. (2007). Use of a Gaussian copula for multivariate extreme value analysis: Some case studies in hydrology. Advances in Water Resources, 30(4):897–912.

[Richardson, 1981] Richardson, C. W. (1981). Stochastic simulation of daily precipitation, temperature, and solar radiation. Water Resources Research, 17(1):182–190.

[Sansom and Thomson, 2010] Sansom, J. and Thomson, P. (2010). A hidden seasonal switching model for high-resolution breakpoint rainfall data. Water Resources Research, 46(8).

[Soubeyroux et al., 2021] Soubeyroux, J.-M., Bernus, S., Corre, L., Drouin, A., Dubuisson, B., Etchevers, P., Gouget, V., Josse, P., Kerdoncuff, M., Samacoits, R., and Tocquer, F. (2021). Les nouvelles projections climatiques de référence DRIAS-2020 pour la Métropole. Technical report, Météo-France.

[Tencaliec et al., 2020] Tencaliec, P., Favre, A.-C., Naveau, P., Prieur, C., and Nicolet, G. (2020). Flexible semiparametric generalized Pareto modeling of the entire range of rainfall amount. Environmetrics, 31(2):e2582.

[Touron, 2019a] Touron, A. (2019a). Consistency of the maximum likelihood estimator in seasonal hidden Markov models. Statistics and Computing, 29(5):1055–1075.

[Touron, 2019b] Touron, A. (2019b). Modélisation multivariée de variables météorologiques. PhD thesis, Université Paris-Saclay (ComUE), https://tel.archives-ouvertes.fr/tel-02319170.

[Vautard et al., 2021] Vautard, R., Kadygrov, N., Iles, C., Boberg, F., Buonomo, E., Bülow, K., Coppola, E., Corre, L., van Meijgaard, E., Nogherotto, R., Sandstad, M., Schwingshackl, C., Somot, S., Aalbers, E., Christensen, O. B., Ciarlo, J. M., Demory, M.-E., Giorgi, F., Jacob, D., Jones, R. G., Keuler, K., Kjellström, E., Lenderink, G., Levavasseur, G., Nikulin, G., Sillmann, J., Solidoro, C., Sørland, S. L., Steger, C., Teichmann, C., Warrach-Sagi, K., and Wulfmeyer, V. (2021). Evaluation of the Large EURO-CORDEX Regional Climate Model Ensemble. Journal of Geophysical Research: Atmospheres, 126(17):e2019JD032344.

[Vidal et al., 2010] Vidal, J.-P., Martin, E., Franchistéguy, L., Baillon, M., and Soubeyroux, J.-M. (2010). A 50-year high-resolution atmospheric reanalysis over France with the Safran system. International Journal of Climatology, 30(11):1627–1644.

[Viterbi, 1967] Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory, 13(2):260–269.

[Vrac et al., 2007] Vrac, M., Stein, M., and Hayhoe, K. (2007). Statistical downscaling of precipitation through nonhomogeneous stochastic weather typing. Climate Research, 34(3):169–184.

[Wilks and Wilby, 1999] Wilks, D. S. and Wilby, R. L. (1999). The weather generation game: a review of stochastic weather models. Progress in physical geography, 23(3):329–357.

[Woollings et al., 2010] Woollings, T., Hannachi, A., Hoskins, B., and Turner, A. (2010). A Regime View of the North Atlantic Oscillation and Its Response to Anthropogenic Forcing. 23(6):1291–1307.

[Yakowitz and Spragins, 1968] Yakowitz, S. J. and Spragins, J. D. (1968). On the Identifiability of Finite Mixtures. The Annals of Mathematical Statistics, 39(1):209–214.

[Zucchini and Guttorp, 1991] Zucchini, W. and Guttorp, P. (1991). A hidden Markov model for space-time precipitation. Water Resources Research, 27(8):1917–1923.