Markov-switching autoregressive models for wind time series.

Pierre Ailliot

Valrie Monbet

Laboratoire de Mathématiques, UMR 6205, Université Européenne de Bretagne, Brest, France

IRMAR, UMR 6625, Université Européenne de Bretagne, Rennes, France

Abstract

In this paper we build a Markov-Switching Autoregressive model to describe a long time series of wind speed measurement. It is shown that the proposed model is able to describe the main characteristics of this time series, and in particular the various time scales which can be observed in the dynamics, from daily to interannual fluctuations.

Keywords: Stochastic weather generators, Wind time series, Markov-switching autoregressive model, Multiscale model, Overdispersion

1 Introduction

This paper develops stochastic models for wind time series over different time scales. A particular impetus for this study was the need to generate realistic wind time series at different meteorological stations located nearby potential wind farms in France, with the aim of assessing various quantities related to the wind power production (see e.g. [9], [11]). However, stochastic models for wind time series have many other risk forecasting applications. For example, they can be used to provide realistic inputs into environmental and ecosystem models (see e.g. [3]), among many other applications (see also [21]).

The most classical approach for modeling wind time series consists in applying the general Box-Jenkins methodology ([7]). Wind times series are generally not stationary, with typically important seasonal and daily components but also interannual variability. The first step of the Box-Jenkins methodology consists in achieving stationarity. Interannual components are generally neglected and various methods are available in the literature for modeling daily and seasonal components, for instance differentiation or scaling with mean and variance functions which evolve periodically (see e.g. [8] and [9]). Another usual approach for treating seasonality in meteorological applications consists in blocking the data by short time periods, typically one month or one season depending on the amount of data available, and then assuming that the changes due to the season can be neglected on this time period. Then, after achieving stationarity, a Box-Cox transformation is generally applied to get a time series with marginal distribution close to a Gaussian distribution and an ARMA model is fitted to this transformed time series (see e.g. [9]).

Box-Jenkins methodology generally leads to a good description of the marginal distribution and second order structure of the original wind time series but fails to reproduce non-linearities which may exist in the dynamics. One well known source of non-linearity in many meteorological time series is induced by the existence of "weather types". They correspond to typical pressure and frontal patterns and induce regime shifts in local weather conditions. For example, for the specific time series considered in this paper, at least two regimes can be easily identified when looking at the data. In the first one, anticyclonic conditions are prevailing leading to steady and low wind speed, whereas in the second one

moving low pressure systems are dominating leading to more important time variability in the wind conditions. This induces heteroscasticity in wind time series, and GARCH models have been proposed in this context (see [25]). In this work, in order to get a more physically-based model, we introduce explicitly the regime shifts through a hidden variable. As concerns meteorological applications, the idea of introducing a latent variable which represents the weather type goes back to [28] where Hidden Markov Models (HMM) were proposed for modelling the space-time evolution of daily rainfall. HMM have then been extensively used for modelling rainfall (see [4] and references therein). HMM are characterized by various conditional independence assumptions on the joint dynamics of the hidden variable and the observations. These assumptions imply in particular that successive observations are conditionally independent given the hidden variable. In practice, it means that all the dynamics should be explained by the weather type, and such assumptions seems rather unrealistic for wind time series since the correlation between successive observations is generally high.

In this paper, we propose to use Markov-Switching AutoRegressive (MS-AR) models. This family of model, which was initially proposed in [14] to describe econometric time series, is a generalization of both HMM and autoregressive models. Indeed, they combine different autoregressive models to describe the evolution of the process at different periods of time, the transition between these different autoregressive models being controlled by a hidden Markov chain like in HMM.

Hereafter, we focus on a particular time series of wind speed measured on the Island of Ouessant ($48^0 27' 36''$ N, $5^0 6' 0''$ W). This time series, which is shown on Figure 1, consists of 51 years of data, from 1948 to 1998, with a data every 6 hours which corresponds to the mean wind speed on a 20 minutes time period. We have chosen to consider a long time series in this study in order to discuss the modelling of interannual variability, but good results were also obtained with the methodology described in this paper on shorter wind time series at various locations in France with different climatologies.

MS-AR models are introduced briefly in Section 2. Then in Section 3, we first block the time series by month, in order to remove the seasonal components, and fit a separate MS-AR model each month. The daily components are included directly in the parametrization

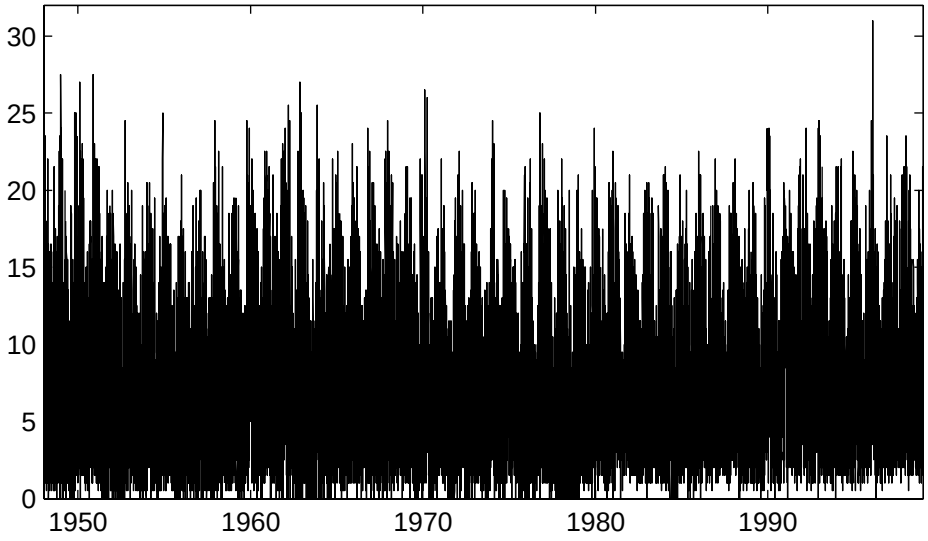

Figure 1: Wind speed in ms^{-1} (y-axis) at Ouessant between 1948 and 1998.

of the MS-AR models. These monthly models are validated by checking their ability to generate realistic wind time series, and it is shown that MS-AR models provide a better description of the short-term dynamics compared to more conventional ARMA models. In Section 4, we explore the seasonality in the parameter values for the monthly models and this leads us to propose an original MS-AR model which incorporates a seasonal component. This new MS-AR is fitted and validated on the entire time series. The results are satisfactory except that the model tends to underestimate the observed interannual variability in the wind condition. This is the well known "overdispersion" phenomenon which is a weakness of many stochastic weather generators (see e.g. [16]). One possible explanation may be the existence of interannual variations which are not taken into account in the model. In Section 5, we thus explore interannual components and show that there is a clear trend in the probability of occurrence of the different weather type. This lead us to propose another MS-AR model which include a trend in the parameter values. It is shown that this new MS-AR permits a better description of the observed interannual variability although it is still underestimated. Finally we conclude in Section 6.

2 MS-AR models

2.1 Model description

A MS-AR process is a discrete-time process with two components $\{S_t, Y_t\}$ where, for our particular application, $\{Y_t\}$ denotes the wind speed process with values in $(0, +\infty)$ and $S_t \in \{1, ..., M\}$ represents the latent weather type at time t. A MS-AR process is then characterized by the two conditional independence assumptions below:

- the conditional distribution of S_t given the values of $\{S_{t'}\}_{t'<t}$ and $\{Y_{t'}\}_{t'<t}$ only depends on the value of S_{t-1}. In other terms, we assume that the weather type $\{S_t\}$ is a first order Markov chain which evolution is independent of the past wind conditions.

- the conditional distribution of Y_t given the values of $\{Y_{t'}\}_{t'<t}$ and $\{S_{t'}\}_{t'\leq t}$ only depends on the values of S_t and $Y_{t-1}, \ldots, Y_{t-p}$. For our particular application, it means that the wind speed process $\{Y_t\}$ is an autoregressive process of order $p \geq 0$ which coefficients evolve in time with the weather type sequence.

When $p = 0$, we retrieve the usual HMM, and the various conditional independence assumptions are summarized by the directed graph below for $p = 1$:

$$\begin{array}{ccccccccc} \cdots & \rightarrow & S_{t-1} & \rightarrow & S_t & \rightarrow & S_{t+1} & \rightarrow & \cdots \\ & & \downarrow & & \downarrow & & \downarrow & & \\ \cdots & \rightarrow & Y_{t-1} & \rightarrow & Y_t & \rightarrow & Y_{t+1} & \rightarrow & \cdots \end{array}$$

MS-AR models were initially introduced in [14] to describe econometric time series, the regimes corresponding to the different states of the economy, and then used for other applications (see e.g. [13] and references therein). MS-AR models for wind time series were initially proposed in [1] and [21] and then, in different contexts, in [3] and [22].

In the usual applications of HMM and MS-AR models, the hidden Markov chain $\{S_t\}$ is supposed to be homogeneous, in which case the transition probabilities $P(S_t = s'|S_{t-1} =$

s) are constant in time and the evolution of $\{S_t\}$ is parametrized by the transition matrix $Q = (q_{s,s'})_{s,s' \in \{1,...,M\}}$ with $q_{s,s'} = P(S_t = s' | S_{t-1} = s)$. HMM with non-homogeneous hidden Markov chain were also proposed for meteorological applications, for example in [20] to describe non-stationary components in time series of wind direction and in [15] to relate the large circulation to local rainfall conditions. In the next sections, we also propose various MS-AR models with non-homogeneous hidden Markov chain in order to describe seasonal and interannual variations.

As concerns the autoregressive models, the most standard MS-AR model is obtained using standard $AR(p)$ models with Gaussian innovations. If $S_t = s_t$, it is assumed that

$$Y_t = a_0^{(s_t)} + a_1^{(s_t)} Y_{t-1} + \ldots + a_p^{(s_t)} Y_{t-p} + \sigma^{(s_t)} \epsilon_t \tag{1}$$

where $\left(a_0^{(s)}, a_1^{(s)}, ..., a_p^{(s)}, \sigma^{(s)}\right) \in \mathbb{R}^{p+1} \times (0, +\infty)$ denotes the unknown parameters of the $AR(p)$ model which describes the evolution of the observed process in the regime $s \in \{1, ..., M\}$ and $\{\epsilon_t\}$ is a sequence of independent and identically distributed Gaussian variable with zero mean and unit variance independent of the Markov chain $\{S_t\}$. In other terms, it is assumed that the conditional distribution $P(Y_t | Y_{t-1} = y_{t-1}, \ldots, Y_{t-p} = y_{t-p}, S_t = s_t)$ is a Gaussian distribution with conditional mean and variance given respectively by

$$E(Y_t | Y_{t-1} = y_{t-1}, \ldots, Y_{t-p} = y_{t-p}, S_t = s_t) = a_0^{(s_t)} + a_1^{(s_t)} y_{t-1} + \ldots + a_p^{(s_t)} y_{t-p} \tag{2}$$

$$var(Y_t | Y_{t-1} = y_{t-1}, \ldots, Y_{t-p} = y_{t-p}, S_t = s_t) = \left(\sigma^{(s_t)}\right)^2 \tag{3}$$

For the application considered in this paper, $\{Y_t\}$ is a process with positive values, and the model with conditional Gaussian distribution may not be appropriate in such situation since it does not permit to restore the constraint $Y_t \geq 0$. In [1], it was proposed to replace the Gaussian distribution by a Gamma distribution and keep (2) and (3) for the conditional moments (the Gamma distribution is also characterized by its two first moments). This model was fitted to various wind time series, and generally good results were obtained (some of these results are reported in [21]). However, one drawback of the parametrization based on the Gamma distribution is that the additional constraints $a_k^{(s)} > 0$ are needed, for $k \in \{0, \ldots, p\}$ and $s \in \{1, \ldots, M\}$, in order to ensure that the

conditional mean of the Gamma distribution is positive. The tests that we have done on various time series indicates that fitting the MS-AR model with Gaussian innovations (1) generally leads to selecting AR models of order $p = 2$ with autoregressive coefficients $a_2^{(s)} < 0$ (see for example the numerical results given in the next sections) and in such situation using the model with conditional Gamma distribution may not be appropriate. Hereafter we consider only models with conditional Gaussian distributions. It permits to save computational time and also avoid numerical problems which may exist when using the Gamma distribution. This and other modelling issues related to the parametrization of the autoregressive models will be further discussed in the next sections.

2.2 Statistical inference

The most classical method for fitting a MS-AR model, for given values of M and p, consists of using the Expectation-Maximization (EM) algorithm. It was proven that this algorithm, which was first introduced in [5] for HMM and then generalized to other models with hidden variables in [12], converges to a maximum of the likelihood function under general conditions (see [27]). The description of the particular form of this algorithm for MS-AR models with Gaussian innovations and homogeneous Markov chain can be found in [19]. This is an iterative procedure, starting from an initial value $\theta(0)$ for the parameters. Each iteration consists of 2 steps:

- **E step**: computation of an auxiliary function $R(\theta, \theta(n))$ which is defined as the conditional expectation of the complete likelihood given the observations and the current value of the parameters. For all the models considered in this paper, this step can easily be performed using the classical forward-backward recursions (see e.g. [10] and [28]).

- **M step**: computation of $\theta(n+1) = argmax_\theta R(\theta, \theta(n))$. Depending on the MS-AR model under consideration, there are not always analytical expression for $\theta(n+1)$, in which case a numerical optimization procedure is required. In order to get an efficient EM algorithm, it is important to implement carefully the optimization problem. In particular, it is often possible to break the optimization problem into

several lower dimensional optimization problems which are much quicker to solve. More precisely, for all the models considered in this paper, it is possible to separate the parameters related to the evolution of the hidden Markov chain, θ_S, and the parameters related to the evolution of the observed process in each regime $s \in \{1, \ldots, M\}$, denoted $\theta_Y^{(s)}$, such that $\theta = \left(\theta_S, \theta_Y^{(1)}, \ldots, \theta_Y^{(M)}\right)$. For example, for the MS-AR model with homogeneous Markov chain and $AR(p)$ models with Gaussian distribution (1), we have $\theta_S = (q_{s,s'})_{s,s' \in \{1,\ldots,M\}}$, with the usual constraints to get a well defined transition matrix, and $\theta_Y^{(s)} = \left(a_0^{(s)}, a_1^{(s)}, \ldots, a_p^{(s)}, \sigma^{(s)}\right)$. Then we have a decomposition of the form

$$R(\theta, \theta(n)) = R_S(\theta_S, \theta(n)) + R_Y^{(1)}(\theta_Y^{(1)}, \theta(n)) + \ldots + R_Y^{(M)}(\theta_Y^{(M)}, \theta(n))$$

which leads to $M + 1$ separate optimization problems on reduced dimension spaces. There may exist analytic expression for some of them, e.g. when the hidden Markov chain is homogeneous or when the autoregressive models are parametrized using (1). Otherwise, a standard quasi-Newton algorithm has been used in this work, with an appropriate treatment of the various constraints on the coefficients.

The EM algorithm has several well-known limitations. First, it may converge to a non-interesting local maximum of the likelihood function. In practice, it means that a careful choice of the starting value has to be made; this is further discussed in the next sections. Another drawback is its slow rate of convergence near the maxima, where it is known that a usual quasi-Newton algorithm is more efficient. An additional advantage of using a quasi-Newton algorithm is that it provides directly an approximation of the Hessian of the log-likelihood function, and thus gives useful information on the variance of the estimates. On the other hand, quasi-Newton algorithms are generally more sensitive to the choice of the starting value and require some programming efforts since they need the gradient of the function to optimize as input to be efficient. Such algorithms have not been implemented in this work.

The stability of MS-AR models and the asymptotic properties of the Maximum Likelihood Estimates (MLE) in HMM and MS-AR models have been studied extensively in the recent years (see e.g. [10] and references therein). In particular, general conditions which ensure

consistency and asymptotic normality of the MLE for MS-AR model with homogeneous hidden Markov chain and the autoregressive models described above with Gaussian or Gamma conditional distributions can be found in [18] and [2], but the existing results do not apply to the non-stationary MS-AR models considered in this paper.

Another important problem in practice, which has received lots of attention in the last few years, is the problem of model selection which aims at finding the "optimal" value of p and M (see e.g. [10] for a recent review). Hereafter, we have chosen to use the Bayes Information Criterion (BIC) as a first guide. Although its use is not justified for MS-AR models from a theoretical point of view, we found that it generally permits to select parsimonious models which fit the data well. It is defined as

$$BIC = -2 \log L + k \log N$$

where L is the likelihood of the data, k is the number of parameters and N is the number of observations. It can be easily computed from the likelihood which is a natural output of the forward recursions performed in the E-step of the EM algorithm.

3 MS-AR model for monthly data

A classical approach for treating seasonality for meteorological time series consists in blocking the data by month and fit a separate model each month, assuming that the different realisations of the same month over different years are independent realizations of a common stochastic process. This approach is used in this section and we discuss the results obtained on the wind time series introduced in Section 1.

3.1 Model description

Even when focusing on a monthly time period, daily fluctuations generally imply that wind time series are not stationary. According to Figure 2, for the time series considered in this work, the wind speed is generally higher during the day than during the night, with maximum mean value at noon and the daily variations are more important in summer

than in winter due to the higher daily variations in the temperature.

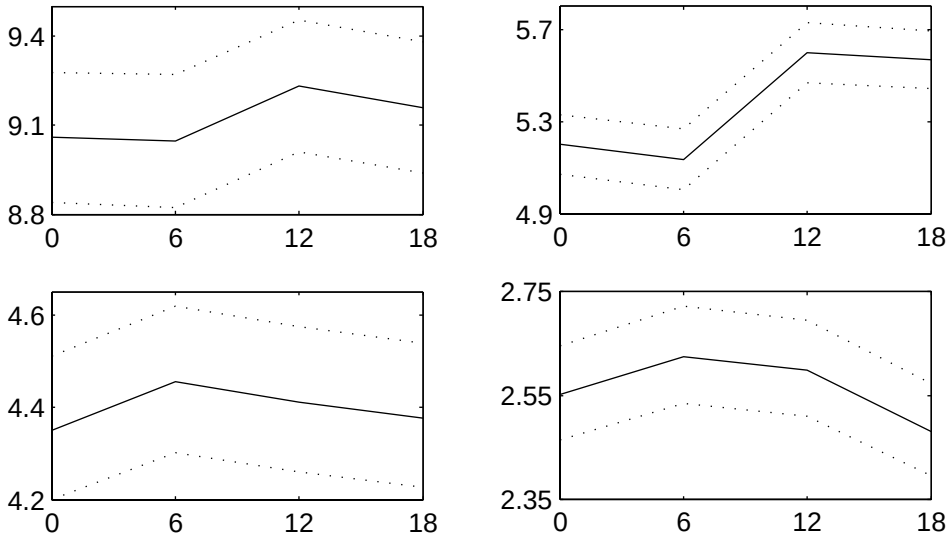

Figure 2: Daily variations for the mean (top) and standard deviation (bottom) of the wind speed in January (left) and July (right). The x-axis represents the time in the day. The dotted lines correspond to 95% confidence interval computed using the (unrealistic) assumption that the observations comes from an i.i.d. Gaussian sample to help interpretation.

A classical approach (see e.g. [9]) for modeling wind time series with daily components consists in scaling the data by subtracting the periodic mean function and eventually dividing by the periodic standard variation function which are shown on Figure 2 and then assume that the residual time series is an $AR(p)$ process. This is equivalent to assume that the wind time series is a non-homogeneous $AR(p)$ process with periodically evolving coefficients. Here we propose to use such non-homogeneous autoregressive models in each regime and replace (2) by (4)

$$E\left(Y_t|S_t = s_t, Y_{t-p} = y_{t-p}, ..., Y_{t-1} = y_{t-1}\right) = a_0^{(s_t)}(t) + a_1^{(s_t)} y_{t-1} + ... + a_p^{(s_t)} y_{t-p} \quad (4)$$

where, for $s \in \{1, ..., M\}$,

$$a_0^{(s)}(t) = \alpha_0^{(s)} + \alpha_1^{(s)} \cos\left(\frac{2\pi}{T_d}(t - \alpha_2^{(s)})\right)$$

with the unknown parameters $\alpha_0^{(s)} \in \mathbb{R}$, $\alpha_1^{(s)} \geq 0$ and $\alpha_2^{(s)} \in [0, 2\pi[$ and T_d represents the number of observations per day in such a way that (4) defines a periodic function with period one day.

In (4), only the intercepts of the AR models are assumed to vary with time. Loosely speaking, it permits to model that the mean of the wind speed exhibits daily variation but not its variance. According to Figure 2, this seems to be realistic for the particular time series considered in this work. This model could be obviously generalized by assuming that the other coefficients of the autoregressive models are periodic functions, but the various attempts that we have done in this direction for our particular data set did not improve the results obtained with the simplest model (4). The coefficient $\alpha_1^{(s)}$ is related to the amplitude of the daily variations in regime s, whereas $\alpha_2^{(s)}$ is associated to the time in the day when the wind speed is maximum. In the limiting case $\alpha_1^{(s)} = 0$ we retrieve the homogeneous model (2). The model (4) allows these characteristics to be different in the weather types. For example, for our dataset we expect more important daily variations when anticyclonic conditions are prevailing than in cyclonic conditions and such behaviour can not be restored by the more conventional approach discussed in [9].

Hereafter, we will also consider another approach which consists in including the daily component in the dynamics of the hidden Markov chain and assume that the transition probabilities are periodic functions. Such approach was initially proposed in [28] for wind direction and then used in [1] for wind speed. In this work, we have considered simple parametric functions and assumed that

$$P(S_t = s' | S_{t-1} = s) \propto q_{s,s'} \exp\left(\kappa_{s'} \cos\left(\frac{2\pi}{T}(t - \phi_{s'})\right)\right) \tag{5}$$

where $Q = (q_{s,s'})_{s,s' \in \{1,...,M\}}$ is a stochastic matrix and, for $s \in \{1, ..., M\}$, $\kappa_s \geq 0$ and $\phi_s \in [0, 2\pi[$ are unknown parameters. Again, $T = T_d$ represents the number of observations per day in such a way that (5) defines a periodic function with period one day. The limiting case $\kappa_s = 0$ for $s \in \{1...M\}$ corresponds to the homogeneous case whereas for high values of κ_s the conditional distribution (5) is concentrated around ϕ_s.

3.2 Parameter estimation

The various models obtained by combining the different parametrizations discussed above for the hidden Markov chain, which can be homogeneous or not, and the autoregressive models which can be homogeneous or not have been fitted to the 12 data sets obtained by blocking the data by month. There are some missing data in the original time series. When only one data is missing, a single linear interpolation method has been used to fill in the gap using adjacent values. It leads to a new time series with missing values only in 1986 and 1991, and this two years include a long time period with no data. We have thus decided to remove these two years from the original time series in order to facilitate the statistical inference. Finally, for each month it remains 49 realizations of length $4 * Nd$, where Nd represents the number of day in the month under consideration, in order to fit and validate the models. These realizations are supposed to be independent and the likelihood function which we consider is obtained as the product of the likelihood over the 49 realizations. The likelihood function has been maximized using the EM algorithm.

In practice, we consider models with a number of regimes $M = 1, \ldots, 5$ and autoregressive models of order $p \in \{1, 2\}$. In order to initialize this algorithm with realistic parameter values, and thus avoid convergence to non-interesting maxima and save computational time, we have used the inclusion of the models. For example, the non-homogeneous models were initialized using the parameter values of the corresponding fitted homogeneous models and the models of order $p = 2$ using the models of order $p = 1$. When such initialization was not available (for example for the homogeneous models of order $p = 1$), the EM has been initialized using several starting values chosen randomly in a set on physically realistic parameter values.

3.3 Results

In this section, we focus on the months of January and July since the results obtained for these two months are representative of the ones for the other months.

According to Table 1, BIC clearly favours MS-AR models with autoregressive models of

order $p = 2$ and a number of regimes M between 2 and 4. In January, BIC selects a model with homogeneous hidden Markov chain and homogeneous autoregressive models, that is a homogeneous model with no daily component. This seems consistent with Figure 2 which suggests that the daily components are not significant in January. In July, when daily components are more important, BIC favours models with homogeneous hidden Markov chain but non-homogeneous autoregressive models. It indicates that it is more appropriate to model daily components inside the dynamics of the weather types than in the dynamics of the weather type. This seems also more natural from a physical point of view since the hidden variable is interpreted as a surrogate of the large scale atmospheric situation which may not be affected by daily components.

January											
	M	1	2	3	4	5	1	2	3	4	5
MC	AR	$p = 1$					$p = 2$				
H	H	29452	28898	28921	28844	28890	29120	**28518**	28524	28546	28577
H	N	29464	28925	28964	29026	28977	29132	28545	28569	28631	28688
N	H	29452	28933	28974	28910	28959	29120	28552	28580	28592	28647
N	N	29464	28960	29014	29094	29028	29132	28580	28619	28700	28768
July											
	M	1	2	3	4	5	1	2	3	4	5
MC	AR	$p = 1$					$p = 2$				
H	H	23572	23299	23295	23367	23408	23380	23100	23109	23190	23233
H	N	23446	23142	23135	23192	23279	23258	22952	**22952**	23028	23110
N	H	23572	23281	23191	23243	23317	23380	23086	23022	23055	23150
N	N	23446	23172	23183	23254	23344	23258	22981	23001	23086	23183

Table 1: BIC values for the various MS-AR models fitted for the months of January and July. The first column indicates if the hidden Markov chain is homogeneous (H) or non-homogeneous (NH), the second column indicates if the AR models are homogeneous (H) or non-homogeneous (NH).

Let us now first focus on January. According to the BIC values given in Table 1, the best model has $M = 2$ regimes, but the difference with the model with $M = 3$ regimes is low. A more precise investigation of these two models shows that the model with $M = 3$

regimes permits to better reproduce some important properties of the data such as the durations of the storms than the model with $M = 2$ regimes. The models with $M \geq 4$ regimes had states with very low probability of occurrence or the fitted states included two very similar states. This led us to restrict attention to $M = 3$ and select the model with homogeneous hidden Markov chain and autoregressive models of order $p = 2$.

According to Table 2, the first regime corresponds to periods with steady wind conditions, with a low standard deviation for the innovation $\sigma^{(s)}$ and also a slower decrease to zero of the autocorrelation functions than in the other regimes, whereas the third regime corresponds to periods with important temporal variability in the wind conditions. The comparison of the means of the stationary distributions in the different regimes also indicates that higher wind speed are generally observed in periods with high variability than in period with low variability. The transition matrix exhibits high values on the diagonal and thus the different regimes are relatively persistent (the mean duration of sojourns varies between 2.69 days in the regime 2 and 5.86 days in regime 3). There are also some very small transition probabilities : for example most of the time the Markov chain will transit from regime 1 to regime 3 through regime 2 and vice-versa. The stationary distribution of the hidden Markov chain indicates that the three regimes have almost the same probability of occurrence.

Transition matrix					AR models				
	S_t				Coefficients				
S_{t-1}	1	2	3	$\pi^{(s)}$	$a_0^{(s)}$	$a_1^{(s)}$	$a_2^{(s)}$	$\sigma^{(s)}$	$\mu^{(s)}$
1	0.92	0.07	0.01	0.35	1.13	0.96	-0.13	1.65	6.63
2	0.07	0.91	0.02	0.35	2.83	0.86	-0.19	2.66	8.77
3	0.01	0.03	0.96	0.30	6.36	0.69	-0.20	3.44	12.32

Table 2: Estimated parameters for the homogeneous model with $M = 3$ regimes and autoregressive models of order $p = 2$ together with the stationary distribution $\pi^{(s)}$ of the Markov chain and the mean $\mu^{(s)}$ of the stationary solution of the AR models. Results for January.

A useful tool to confirm visually the interpretation of the states consists in computing the

smoothing probabilities defined as the conditional distribution of the hidden state given all the observations $(y_1, ..., y_N)$ available in a given month

$$P[S_t = s|Y_1 = y_1, ..., Y_N = y_N]$$

for $s \in \{1, ..., M\}$. The smoothing probabilities can be used to compute the "most likely regime" at each time step and then segment the observed time series according to the different regimes. An example of such segmentation is shown on Figure 3 : we retrieve periods with low variability and periods with more variability.

Figure 4 shows the distribution of the wind direction in the different regimes identified using the smoothing probabilities. The third regime is mainly associated with wind from the South-West : it may correspond to cyclonic conditions when quickly evolving low-pressure systems are coming from the Atlantic ocean. The two other regimes have similar distributions for the wind direction and can be associated to all wind direction. Looking at the distribution of other meteorological variables, such as the sea-level pressure, may help refining the meteorological interpretation of the different regimes.

To further validate the model, we have checked its ability to simulate realistic wind time series since this is an important aspect for the applications which have motivated this work. For that, we have generated artificial time series from the model and we have compared various statistics computed from these artificial sequences with those computed from the data. Typical results are shown in Figure 5. In order to assist visual comparison, 95% prediction intervals for the fitted model have been superimposed where these quantities have been computed using Monte Carlo methods. The limits of the intervals correspond to the 2.5th and 97.5th percentiles from 1000 independently sequences of 49 months of January simulated using the fitted model.

Figure 5 shows that the results obtained for the distribution function of the marginal distributions, the autocorrelation function and the distribution function of the sojourn durations above and below some selected thresholds. These results were compared with the ones obtained using the Box-Jenkins methodology, and we could identify several advantages of using MS-AR models. First, the MS-AR model is able to reproduce the marginal distribution of the process without applying an initial transformation, such as

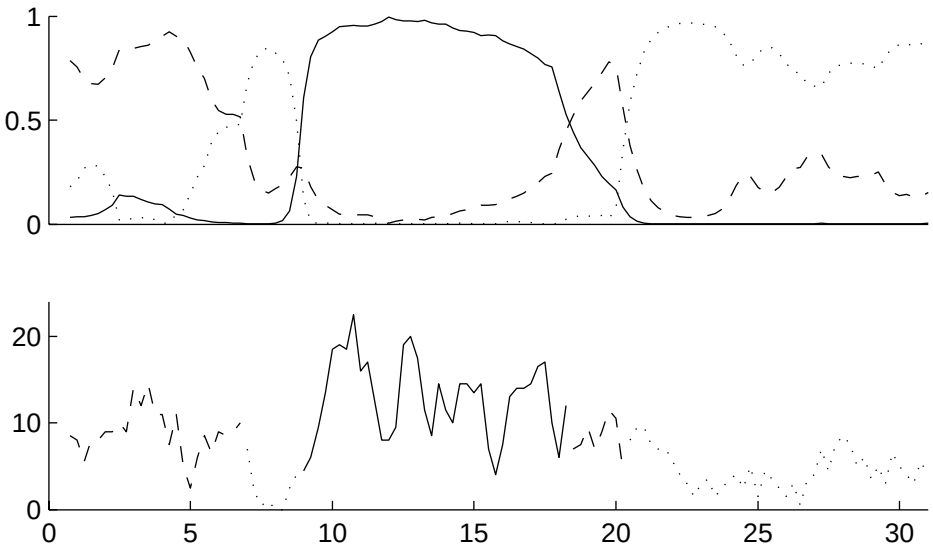

Figure 3: Top panel : smoothing probabilities $P[S_t = s|Y_1 = y_1, ..., Y_T = y_T]$ for $s = 1$ (dotted line), $s = 2$ (dashed line) and $s = 3$ (full line) for one month of January. Bottom panel : wind speed $(y_{1-p}, ..., y_T)$ for the same month of January. The line style corresponds to the most likely state with the same convention than on the top panel.

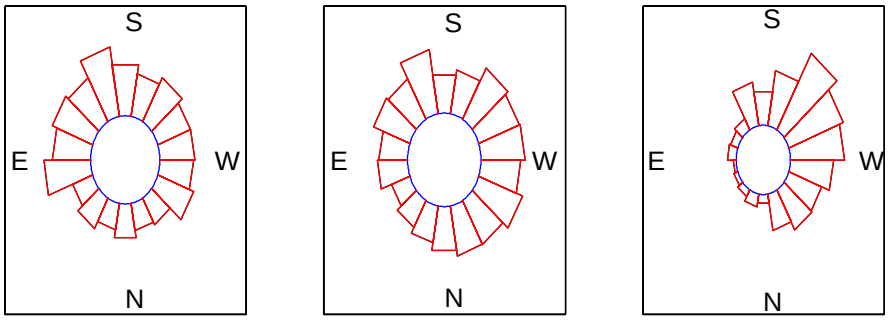

Figure 4: Wind direction in the different regimes identified by the smoothing probabilities. Results for January.

the Box-Cox transformation, to achieve normality. This is not surprising given the distributional flexibility inherent in hidden Markov modelling. However, the model tends to overestimate the probability of low wind speed and can even simulate negative wind speed. Nevertheless, the results remains satisfactory, especially if the simulated time series are used as input to simulate the behaviour of a system with a low sensitivity to light wind

conditions, such as the power value of a wind turbine. If low wind speed are important for a particular application, using the model with conditional Gamma distributions discussed in Section 2 could be more appropriate. Figure 5 shows that the fitted MS-AR model permits also to reproduce the autocorrelation function and the distribution functions of the sojourn durations above and below some selected thresholds since the sample distribution function always lie in the 95% prediction interval computed from the model. We could not get such good results using Box-Jenkings methodology as concerns the sojourn durations and there are good theoretical reasons for that. Indeed, Box-Jenkins methodology is based on ARMA models and thus assume that, after eventual increasing transformation, the process is Gaussian. It entails some symmetry in the dynamics of the time series and that the behaviour for low wind speed should be similar to the one for high wind speed. In particular, the durations of the sojourns below the quantile of order p should have the same distribution that the ones above the quantile of order $1-p$. Figure 5 indicates that this is not true for the time series considered in this paper and that the durations of the excursions below the 25% quantile tend to be longer that the ones above the 75% quantile : we find again that the time series exibits more variability at high level than at low level. MS-AR models mix different AR models and thus allow, for example, different dynamics at low and high levels. It leads to a good reproduction of the sojourn durations (see Figure 5).

Similar results were obtained for other months and at other locations. For the months with important daily variations, we also checked that the fitted MS-AR model with homogeneous hidden Markov chain but non-homogeneous AR models can reproduce the characteristics of these variations. For example, Figure 6 shows that the fitted model can reproduce both the fluctuations of the mean wind speed and the peak at $1day^{-1}$ in the periodogram for the month of July.

4 MS-AR model with seasonal components

In the previous section, a separate MS-AR model was fitted each month. For many applications, it is necessary to have a model which can simulate the wind speed on a

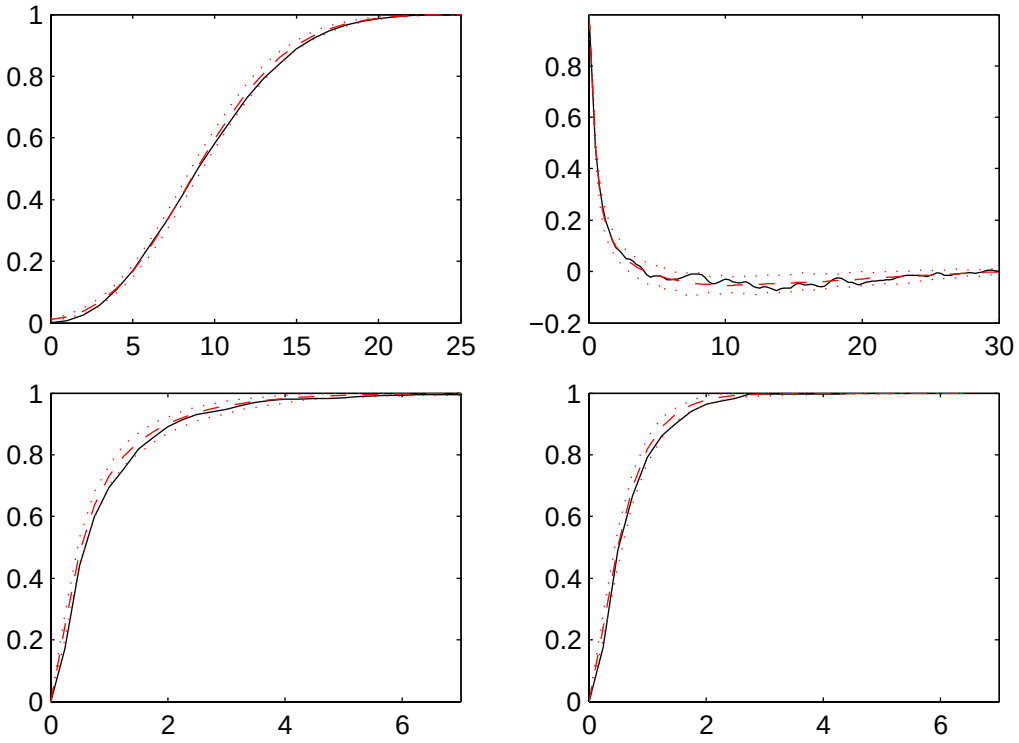

Figure 5: Top left: cumulative distribution function of the marginal distribution. Top right : autocorrelation function. Bottom left: cumulative distribution function of the time duration of the sojourns below the threshold $6ms^{-1}$ which corresponds to the 25% quantile of the marginal distribution. Bottom right: cumulative distribution function of the time duration of the sojourns above $12ms^{-1}$ which corresponds to 75% quantile of the marginal distribution. Time is expressed in days. The full line corresponds to the sample functions and the dashed line to the fitted model with a 95% prediction intervals (dotted line). The distributions for the fitted model was obtained by simulation. Results for January.

yearly basis. A straightforward combination of the monthly models would lead to a yearly model where the parameters vary as a step function with a break at the beginning of each month. In this section, we propose including seasonality in a more appropriate way into the model.

The results obtained when fitting the MS-AR models introduced in the previous section to each of the 12 months indicate that a MS-AR with $M = 3$ regimes, homogeneous hidden Markov chain but non-homogeneous autoregressive models is the simplest model which

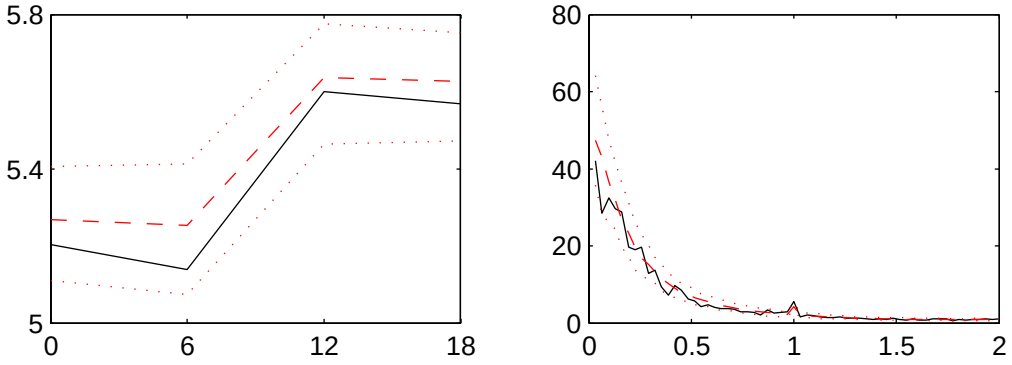

Figure 6: Left panel : daily variations of the mean wind speed. Right panel : periodogram (on x-axis in day^{-1}) . The full line corresponds to the sample functions and the dashed line to the fitted model with a 95% prediction intervals dotted line). The distribution for the fitted model was obtained by simulation. Results for July.

gives satisfactory results for all months. As discussed in Section 3, simplest homogeneous MS-AR models also provide a good description of wind conditions in winter, when daily components can be neglected, but for simplicity reasons we have decided to keep the same model for the different months. Then, in order to be able to follow the seasonal evolution of the parameters, the regimes have been numbered increasingly according to their conditional standard deviations $\sigma^{(s)}$, the first regime corresponding to wind conditions with low variability whereas the third one to higher variability. In May the first two states were inverted in order to make the time evolution of the coefficients more consistent.

Figure 7 provides a synthetic view of the seasonal evolution of some of the coefficients of the fitted models and summarizes important features of the climatology. First, the time evolution of $\alpha_0^{(s)}$ and $\sigma^{(s)}$ indicates respectively that the mean and the temporal variability are generally higher in winter than in summer. Then, the amplitude of the daily component $\alpha_1^{(s)}$ is maximum in spring and summer in regime 1 and 2 and at the end of summer in regime 3. The comparison of the values of $\alpha_1^{(s)}$ and $\alpha_0^{(s)}$ in the different regimes $s \in \{1, ..., M\}$ shows that the contribution of the daily component to the mean wind speed is more important in the regimes with low temporal variability. Finally, the diagonal coefficients of the transition matrices indicate that the third regime is more persistent in winter than in summer, and thus cyclonic conditions may generally last

longer in winter than in summer.

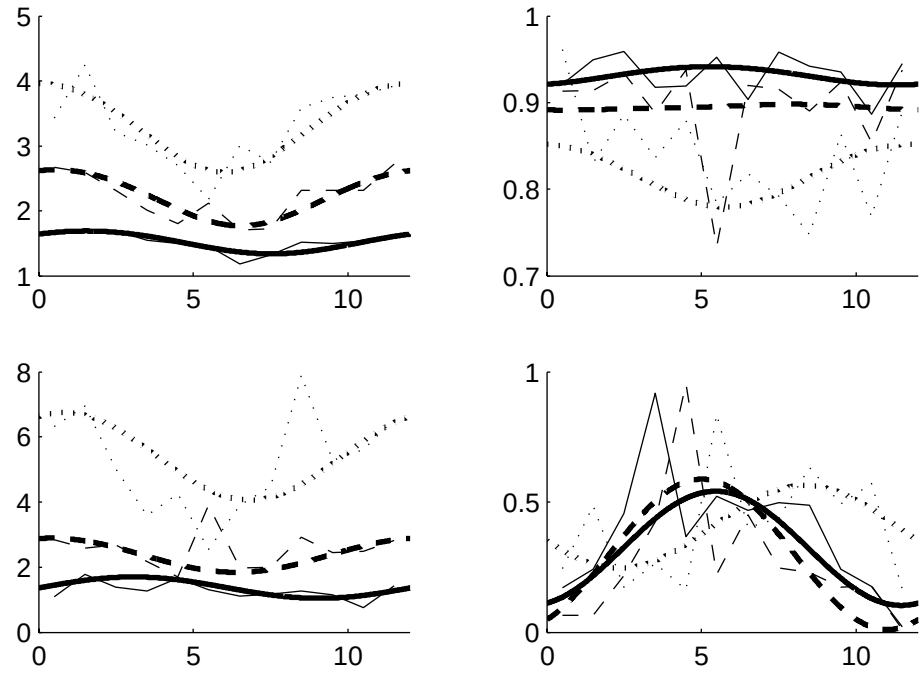

Figure 7: Seasonal variations of $\sigma^{(s)}$ (top left), $q_{s,s}$ (top right), $\alpha_0^{(s)}$ (bottom left) and $\alpha_1^{(s)}$ (bottom left). The full line corresponds to regime $s = 1$, the dashed line to $s = 2$ and the dotted line to $s = 3$. The thin line corresponds to the values obtained when fitting separately the models to monthly data whereas the thick line corresponds to the values obtained after fitting the seasonal model on yearly data. The x-axis represents the time in month.

Figure 7 also suggests to let the coefficients of the model evolve smoothly in time instead of using step functions. In this work, we use simple parametric forms to describe the seasonal evolution of the different coefficients. More precisely, we used again (5) for the transition matrix but with $T = T_y$ the number of observations in one year in order to obtain a periodic function with period one year. Then we use AR models of order $p = 2$ with time varying coefficients to model the conditional evolution of the wind speed

$$Y_t = a_0^{(s_t)}(t) + a_1^{(s_t)}(t)Y_{t-1} + a_2^{(s_t)}(t)Y_{t-p} + \sigma^{(s_t)}(t)\epsilon_t \tag{6}$$

with $\{\epsilon_t\}$ a sequence of independent and identically distributed Gaussian variable with zero mean and unit variance independent of the Markov chain $\{S_t\}$. Again the daily

component is modelled assuming that

$$a_0^{(s)}(t) = \alpha_0^{(s)}(t) + \alpha_1^{(s)}(t)\cos\left(\frac{2\pi}{T_d}(t - \alpha_2^{(s)}(t))\right)$$

Then, if $f(t)$ denotes the value of one of the parameters of the AR models at time t (i.e. $f(t) = a_1^{(s)}(t)$, $f(t) = a_2^{(s)}(t)$, $f(t) = \sigma^{(s)}(t)$, $f(t) = \alpha_0^{(s)}(t)$, $f(t) = \alpha_1^{(s)}(t)$ or $f(t) = \alpha_2^{(s)}(t)$ for some $s \in \{1, ..., M\}$), we assume a smooth seasonal evolution of the form

$$f(t) = f_0 + f_1 cos\left(\frac{2\pi}{T_y}(t - f_2)\right) \tag{7}$$

with f_0, $f_1 \geq 0$ and $f_2 \in [0, 2\pi[$ unknown parameters. Since the conditional standard deviation $\sigma^{(s)}(t)$ and the amplitude of the daily component $\alpha_0^{(s)}(t)$ should be positive in order to ensure identifiability, the constraints $f_0 > 0$ and $f_1 < f_0$ were added when $f(t) = \sigma^{(s)}(t)$ or $f(t) = \alpha_0^{(s)}(t)$ for $s \in \{1, ..., M\}$.

Due to the complexity of the model and the length of the time-series under consideration, it is important to initialize the EM algorithm with realistic parameter values. Indeed, each iteration of the EM algorithm requires important CPU time and thus the number of iteration needs to be reasonable. Furthermore, using arbitrary values is very likely to lead the algorithm to converge to a non-interesting maximum of the likelihood function. In practice, the parameters have first been estimated using the least square method and the parameter values obtained when fitting separately the models to each month of data and then reestimated using the EM algorithm on the whole time series.

We obtain a non-stationary model, which includes both daily and seasonal components and which can be used to generate long wind time series with no discontinuity problems at the beginning of each month. Again, like in Section 3.3, the realism of the simulated time series has been checked by comparing various statistics computed from the synthetic time series to the ones of the original data. We first performed validation on a monthly basis (we considered both calendar month and also periods from the 15th of one month to the 15th of the following month), and the results were similar to those reported in Section 3.3. This is not surprising since, according to Figure 7, the restriction of the fitted seasonal model to a monthly time period is very close to the models fitted on monthly data.

We also performed validation on a yearly basis, and in particular we checked the ability of the model to reproduce the interannual variability of the wind conditions. Figure 8

shows that the fitted model underestimates the observed variability in the yearly mean and yearly maximum wind speed. Monthly or seasonal validation leads to similar results. This is a well known feature of many stochastic weather generators which is termed "overdispersion" in the literature (see e.g. [16] and [17]). Two possible sources of overdispersion are identified in [16]. The first one is an inadequate modelling of high-frequency variations and the second one is the presence of low-frequency variations in the climate, on an interannual time scale, which are not taken into account by the model (see also [6]). The results given in Section 3.3 indicate that the model is able to reproduce the short-term dynamics and in particular the autocorrelation function up to time lags of one month (see Figure 5). As a consequence, in absence of interannual components, the variability of the monthly mean should also be well described by the model since for a second-order stationary process the variance of the sample mean can be deduced from the autocovariance function. In the next section we thus investigate the presence of interannual components in the time series under consideration and show that including it into the model help reproducing the interannual variability.

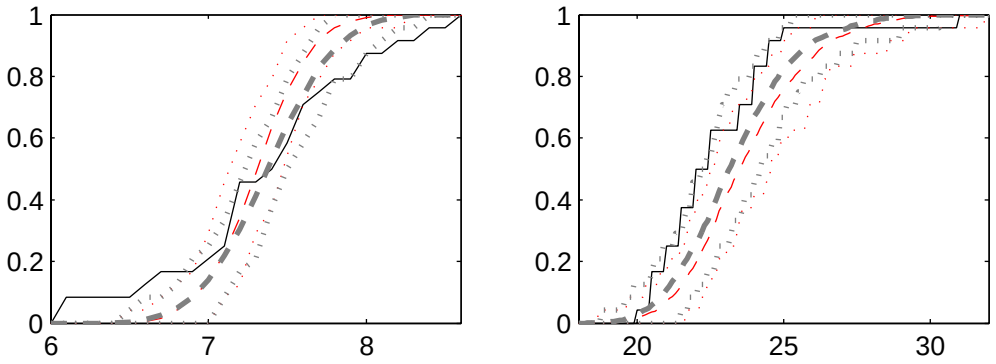

Figure 8: Distribution function of the annual mean (left) and annual maxima (right). The full line corresponds to the sample function and the dashed line to the fitted models with a 95% prediction intervals (dotted line). The thin lines correspond to the seasonnal model without trend (see Section 4) and the thick lines to the model with trend (see Section 5). The distribution for the fitted model was obtained by simulation. Results for the time period 1973-1998.

5 Model with interannual components

Figure 9 shows the 7-year running mean of the conditional expectation $E[S_t|Y_1 = y_1, ..., Y_T = y_T]$ associated to the smoothing probabilities for the seasonal model introduced in Section 4. This time series exhibits a clear trend, with low values in the years 1950-1955 and 1970-1975, high values for the years 1960-1965 and a tendency to increase from 1970 to 2000. According to the interpretation of the various states given in Section 3.3, higher expectations may correspond to periods with higher temporal variability in the wind conditions and thus to periods with more frequent cyclonic conditions.

It is well known that many long meteorological time series exhibit non-climatic (or artificial) sudden changes due, for example, to an instrument change or a change in station location or exposure and that this may affect the study of the climatic trends. Since we have few information on the existence of such changes for the time series considered in this paper or access to an homogenized version of this time series, the results have been compared with those obtained using reanalysis data. More precisely, we have used the ERA-40 data set which consists in a global reanalysis with 6-hourly data covering the period from 1958 to 2001. This reanalysis was carried out by the European Centre for Medium Range Weather Forecast (ECMWF) and can be freely downloaded and used for scientific purposes at the URL:
http://data.ecmwf.int/data

The seasonal model described in Section 4 was then fitted to the time series retrieved from the ERA-40 data set for the same location than in situ-data. We obtained generally similar estimation for the parameters of the model, except for the conditional standard deviations $\sigma^{(s)}$ which are systematically lower for ERA-40 time series: reanalysis data tend to be smoother than in-situ data. Figure 9 also shows the 7-year running mean of the conditional expectation associated to the smoothing probabilities computed using ERA-40 data. We also observe a clear trend and comparing with in-situ data indicates a good overall agreement. For comparison purpose, Figure 9 also shows the Atlantic multidecadal oscillation (AMO) index which is significantly positively correlated with the running means of the smoothing expectation shown on the same figure: periods

with higher values of the AMO index seems to coincide with less frequent steady wind conditions. This may be an indication that the observed trend may partly be explained by climatic variations. However, there are also differences between the results obtained using ERA-40 and in-situ data which may correspond to non-climatic breaks in one of these two time series (see also [23]).

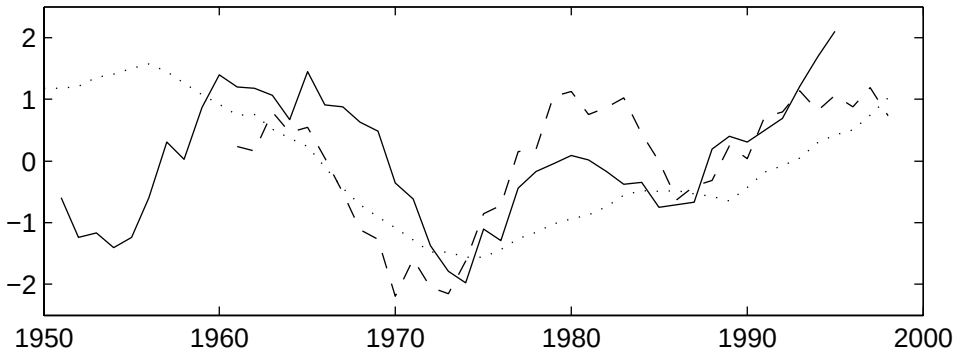

Figure 9: 7-year running mean of the smoothing expectation $E[S_t|Y_{1-p} = y_{1-p}, ..., Y_T = y_T]$ for the seasonal model together with the AMO index (dotted line). The full line corresponds to the model fitted on in-situ data whereas the dashed line corresponds to the model fitted on ERA40 data at the same location. The three time series have been scaled by removing the mean and dividing by the standard deviation in order to facilitate the comparison.

In order to study the impact of the interannual variations on the overdispersion, we have chosen to focus on the time period from 1973 to 1998 when the running mean for in-situ data shown on Figure 9 exhibits a clear increasing trend, and replaced (5) by

$$P(S_t = s'|S_{t-1} = s) \propto q_{s,s'} \exp\left(\kappa_{s'} \cos\left(\frac{2\pi}{T_y}(t - \phi_{s'})\right) + \lambda_{s'} t\right) \tag{8}$$

where, for $s \in \{1...M\}$, λ_s is an unknown parameter which describe possible trends in the probability of occurrence of the regime s. Here, we assume that the long-term climatic variations only impact the probability of occurrence of the different weather types but not the dynamics inside the weather types. Non-homogeneous hidden Markov models, based on similar conditional independence assumptions, have already been proposed in

the literature for statistical downscaling (see e.g. [15] and [26]), in which case the hidden weather type is used to link the large-scale circulation to local weather condition.

Again, the model has been fitted using the EM algorithm. Figure 10 indicates that the fitted model is able to reproduce the observed trend in the probability of occurrence of the different weather types. The estimation for the parameter λ_1 is negative, and it coincides with the fact that regime 1 is less and less likely whereas positive values for λ_2 and λ_3 indicate that the two regimes with more variability become more and more likely.

Figure 8 shows that the model with interannual components better reproduces the variability of the observed mean and maximum values compared to the model without interannual component, but sill underestimates the observed variability.

Using a more sophisticated model for the interannual components could again improve these results. For example, we could replace the linear trend in (8) by a polynomial function, include covariates such as the AMO index, or consider models where the interannual components also modify the dynamics inside the regimes. An alternative in order to improve the description of the interannual variability could consist in using more sophisticated models for the seasonal component (see e.g. [24]).

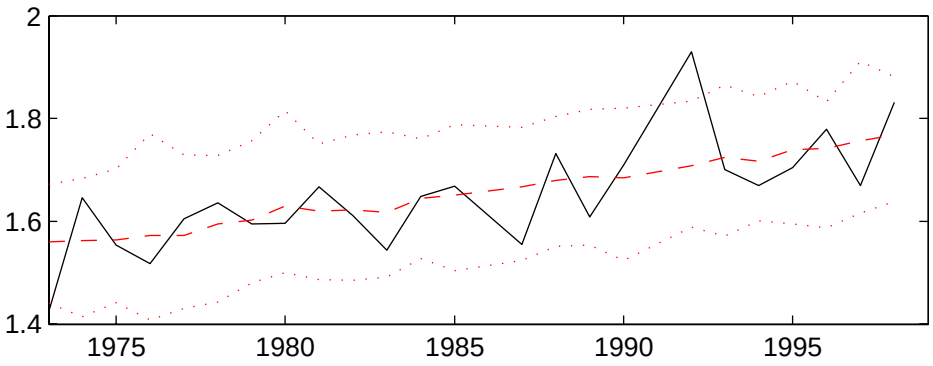

Figure 10: Annual mean of the smoothing expectations $E[S_t|Y_1 = y_1, ..., Y_T = y_T]$ for the seasonal model with interannual component (full line) and annual mean of the expectation of the Markov chain with transition probabilities (8) (dashed line) with a 95% prediction intervals (red dotted line). The expectation for the fitted model was obtained by simulation. Results for the time period 1973-1998

6 Conclusions

This paper investigates the use of MS-AR models to describe wind time series and it is shown that these models have several virtues. First, thanks to their distributional versatility, they are able to describe the marginal distribution of the time series and thus pre-processing the data, like applying the Box-Cox transformation, is not needed. Then, these models have the ability to model diverse time scales which are present in wind time series and improve the description of important properties of the dynamics such as the durations of calm or stormy conditions. This is an important aspect for many applications of these models. Finally, their interpretability leads to open structure which allows for more physical models. In this work, this is used to include various time scale, from daily to interannual components, in a realistic manner into the model.

References

[1] P Ailliot. *Modéles autorégréssifs à changements de régimes makoviens. Application aux séries temporelles de vent.* PhD thesis, Université de Rennes 1, 2004.

[2] P. Ailliot. Some theoretical results on a markov-switching autoregressive models with gamma innovations. *Comptes Rendus de l'Acadmie des Sciences de Paris*, 343(4):271–274, 2006.

[3] P. Ailliot, E. Frénod, and V. Monbet. Long term object drift forecast in the ocean with tide and wind. *Multiscale Modeling and Simulation*, 5(2):514–531, 2006.

[4] P. Ailliot, C. Thompson, and P. Thomson. Space time modeling of precipitation using a hidden markov model and censored gaussian distributions. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 58(3):405–426, 2009.

[5] L.E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *Annals of Mathematical Statistics*, 41:164–171, 1970.

[6] J.C. Bouette, J.F. Chassagneux, D. Sibai, R. Terron, and A. Charpentier. Wind in ireland: long memory or seasonal effect? *Stochastic Environmental Research and Risk Assessment*, 20(3):141–151, 2006.

[7] G.E.P. Box and G.M. Jenkins. *Time series analysis, forecasting and control (revised edn.)*. Holden-Day, San Francisco., 1976.

[8] P.J. Brockwell and R.A. Davis. *Introduction to Time Series and Forecasting, second edition*. Springer-Verlag, New York, 2002.

[9] B.G. Brown, R.W. Katz, and A.H. Murphy. Time series models to simulate and forecast wind speed and wind power. *Journal of climate and applied meteorology*, 23:1184–1195, 1984.

[10] O. Cappé, E. Moulines, and Rydén T. *Inference in hidden Markov models*. Springer-Verlag, New York, 2005.

[11] F. Castino, R. Festa, and C.F. Ratto. Stochastic modelling of wind velocities time series. *Journal of Wind Engineering and industrial aerodynamics*, 74:141–151, 1998.

[12] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[13] Y. Ephraim and N. Merhav. Hidden markov processes. *IEEE Transactions on Information Theory*, 48:1518–1569, 2002.

[14] J.D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57:357–384, 1989.

[15] J.P Hughes and P. Guttorp. A class of stochastic models for relating synoptic atmospheric patterns to local hydrologic phenomenon. *Water Resources Research*, 30:1535–1546, 1994.

[16] R.W. Katz and M.B. Parlange. Overdispersion phenomenon in stochastic modeling of precipitation. *Journal of Climate*, 11:591601, 1999.

[17] R.W. Katz and X. Zheng. Mixture model for overdispersion of precipitation. *Journal of Climate*, 12:2528–2537, 1999.

[18] V. Krishnamurthy and T. Ryden. Consistent estimation of linear and non-linear autoregressive models with markov regime. *Journal of time series analysis*, 19(3):291–307, 1998.

[19] H.M. Krolzig. *Markov-switching vector Autoregressions. Modelling, statistical inference and applications to business cycle analysis*. Lecture notes in economics and mathematical systems, Springer-Verlag, Berlin, 1997.

[20] I.L. McDonald and W. Zucchini. *Hidden Markov and Other Models for Discrete-Valued Time Series*. Chapman & Hall/CRC, London, 1997.

[21] V. Monbet, P. Ailliot, and M. Prevosto. Survey of stochastic models for wind and sea-state time series. *Probabilistic Engineering Mechanics*, 22(2):113–126, 2007.

[22] P. Pinson, L.E.A. Christensen, H. Madsen, P.E. Sorensen, M.H. Donovan, and Jensen L.E. Regime-switching modelling of the fluctuations of offshore wind generation. *Journal of Wind Engineering and Industrial Aerodynamics*, 96(12):2327–2347, 2008.

[23] P.A. Pirazzoli, H. Regnauld, and L. Lemasson. Changes in storminess and surges in western france during the last century. *Marine Geology*, 210:307–323, 2004.

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

[25] R.S.J. Toll. Autoregressive conditional heteroscedasticity in daily wind speed measurements. *Theoretical and Applied Climatology*, 56:113–122, 1997.

[26] M. Vrac, M. Stein, and K. Hayhoe. Statistical downscaling of precipitation through nonhomogeneous stochastic weather typing. *Climate Research*, 34:169–184, 2007.

[27] C.F.J. Wu. On the convergence properties of the em algorithm. *Annals of Statistics*, 11(1):95–103, 1983.

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