



HAL
open science

Non-homogeneous hidden Markov-switching models for wind time series

Pierre Ailliot, Julie Bessac, Valérie Monbet, Françoise Pene

► **To cite this version:**

Pierre Ailliot, Julie Bessac, Valérie Monbet, Françoise Pene. Non-homogeneous hidden Markov-switching models for wind time series. *Journal of Statistical Planning and Inference*, 2015, 160, pp.75-88. 10.1016/j.jspi.2014.12.005 . hal-00974716v2

HAL Id: hal-00974716

<https://hal.science/hal-00974716v2>

Submitted on 19 Jun 2014

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

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

Non-homogeneous hidden Markov-switching models for wind time series

Pierre Ailliot¹, Julie Bessac², Valérie Monbet², Françoise Pène¹

June 19, 2014

¹ *Laboratoire de Mathématiques de Bretagne Atlantique, UMR 6205, Université de Brest, France*

² *Institut de Recherche Mathématiques de Rennes, UMR 6625, Université de Rennes 1, France*

June 19, 2014

Abstract

In this paper we propose various Markov-switching autoregressive models for bivariate time series which describe wind conditions at a single location. The main originality of the proposed models is that the hidden Markov chain is not homogeneous, its evolution depending on the past wind conditions. It is shown that they have good theoretical properties and permit to reproduce complex features of wind time series such as non-linear dynamics and the multimodal marginal distributions.

Keywords: Wind time series, Markov-switching autoregressive process, non-homogeneous hidden Markov process, linear-circular time series, consistency

1 Introduction

Meteorological time series are a key input in many risk forecasting and impact studies applications and historical data are often available over periods of time that are not long enough to get reliable estimates of the quantities of interest. Stochastic weather generators have been developed to overcome this insufficiency by simulating artificial sequences of unlimited number of meteorological variables with statistical properties similar to those of the observations (see [22] and references therein). These generators can also be useful for downscaling global climate models (see e.g. [18] and references therein) or infilling missing values by conditional simulation. In this paper we focus on wind time series. Various approaches have been proposed in the literature for modeling time series of wind speed (see e.g. [20] for a review). In comparison, there exists only very few models for circular time series of wind direction or for bivariate time series describing simultaneously the evolution of the wind speed and the wind direction. There is thus a need for models which can reproduce the specificities of such time series and this paper aims at filling this gap.

In this work we consider a wind time series for a location off the French Brittany coast which bivariate marginal distribution has complex features (see Figure 1). In particular it clearly exhibits two modes, each one corresponding to a different meteorological regime or 'weather type':

the prevailing mode (westerly winds) corresponds to cyclonic conditions, i.e. low pressure systems (e.g. storms) coming from the North-Atlantic ocean, whereas the second mode (easterly winds) corresponds to anticyclonic conditions which can temporarily deviate or block the westerly flow. The alternation of such weather regimes is a well-known characteristic of the North-Atlantic/European area, and after [24] brought some evidence for quasi-stationary solutions in the equations of the atmospheric flow, thus giving a physical meaning to the statistically-derived regimes, they have been broadly used in climate studies. More generally, the presence of regimes with distinct weather conditions is a usual feature of meteorological time series and a classical approach for modeling these meteorological regimes consists in introducing a hidden (or latent) variable. This idea goes back to [26] where Hidden Markov Models (HMMs) were proposed for modeling the space-time evolution of daily rainfall. HMMs have also been proposed for modeling time series of wind direction in [27]. However HMMs assume that the successive observations are conditionally independent given the latent weather type and cannot the strong relationship which exists between the wind conditions at successive time steps for the dataset considered in this work. Markov-Switching AutoRegressive (MS-AR) have been proposed in this context to model time series of wind speed in [20, 21, 1]. MS-AR models extend both the usual HMMs, by adding dynamics in the regimes, and AR models, which are often used to model wind time series (see e.g. [6]), by introducing regime switchings through a latent variable.

In HMMs or MS-AR models, the evolution of the weather type is independent of the past weather conditions. For our particular example, it would imply for example that the probability of switching from the cyclonic conditions to the anticyclonic conditions between time t and time $t + 1$ does not depend on the wind conditions observed at time t whereas we know that these switchings generally occur when the wind is blowing from the north and is very unlikely to occur when the wind is blowing from the south. One originality of the models proposed in this paper is that the evolution of the latent weather type depends on past wind direction leading to non-homogeneous MS-AR (NHMS-AR) models. We show that NHMS-AR models lead to a better description of important characteristics of the data considered in this work, such as multimodality and non-linear dynamics, compared to MS-AR models.

The wind condition at a single location at time t can be described using the polar coordinates $\{U_t, \Phi_t\}$, where U_t denotes the wind speed with values in \mathbb{R}^+ and Φ_t the wind direction with values in $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$ or the Cartesian coordinates $\{u_t, v_t\}$ where u_t and v_t denote respectively the zonal and meridional components with values in \mathbb{R} . The polar coordinates are generally used by meteorologists, probably because they are easier to interpret. However, from a statistical point of view, it is probably more straightforward to model the time series of Cartesian components since many models, such as Gaussian vector AR models, have been proposed for bivariate time series with values in \mathbb{R}^2 whereas the process $\{U_t, \Phi_t\}$ is a linear-circular process with values in $\mathbb{R}^+ \times \mathbb{T}$ and very few models have been proposed for such variables. Both representations are considered in this paper and a discussion of their respective advantages is given.

The paper is organized as follows. NHMS-AR models are introduced in Section 2 with specific parameterizations proposed when considering Cartesian and polar coordinates. In Section 3, we briefly describe the EM algorithm which has been used to maximize the likelihood function and discuss the asymptotic properties of the maximum likelihood estimates (MLE). Then the performances of the models are discussed and compared in Section 4. The data used in this paper are also introduced at the beginning of this section. At last, we make a synthesis of the obtained results and we give some perspectives in Section 5.

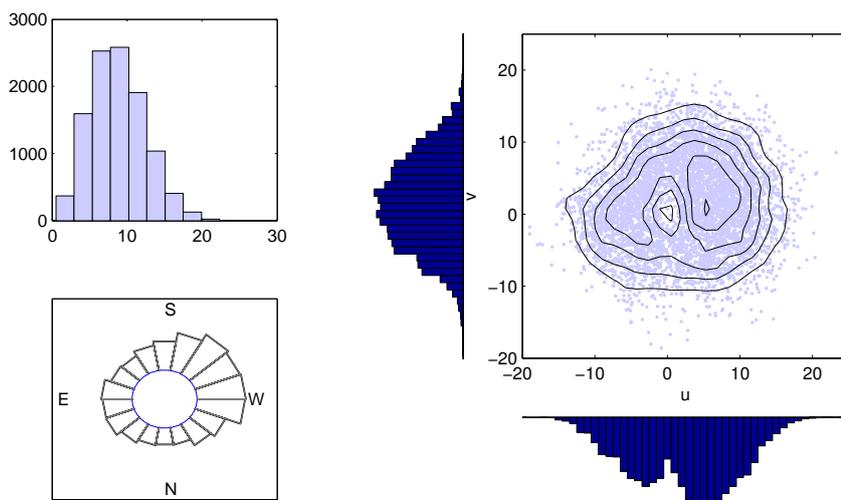


Figure 1: Histogram of $\{U_t\}$ (top left), rose plot of $\{\Phi_t\}$ (bottom left) and histograms of $\{u_t\}$ and $\{v_t\}$ and joint distribution of $\{u_t, v_t\}$ (right). The lines on the scatter plots are levels of a non-parametric kernel estimate of the bivariate density. Results for the months of January.

2 Models

2.1 Non-homogeneous Markov-switching autoregressive models

Let $X_t \in \{1, \dots, M\}$ represents the latent weather type and Y_t denotes the observed wind conditions at time t . Throughout the article $\{Y_t\}$ will represent successively the bivariate process of Cartesian coordinates of wind in Section 2.3, the wind direction in Section 2.4, and finally $\{Y_t\}$ stands for the wind speed in Section 2.5. Let us write \mathbb{E} for the space in which Y_t takes values (\mathbb{E} will respectively refer to \mathbb{R}^2 , \mathbb{T} and \mathbb{R}^+ in the following sections). It will be useful to introduce notation $Y_t^{t+u} := (Y_t, \dots, Y_{t+u})$, $y_t^{t+u} := (y_t, \dots, y_{t+u})$ (as well as X_t^{t+u} , x_t^{t+u}) for $t > 0$ and $u > 0$.

Hypothesis 1. *Let $s, M \geq 1$ be some integers. The sequence $(X_t, Y_{t-s+1}^t)_{t \in \mathbb{Z}}$ follows a NHMS-AR model if it is a Markov chain with values in $\{1, \dots, M\} \times \mathbb{E}$ such that*

- *the conditional distribution of X_t given the values of $\{X_{t'}\}_{t' < t}$ and $\{Y_{t'}\}_{t' < t}$ only depends on X_{t-1} and Y_{t-1} and we denote $p_1(x_t | x_{t-1}, y_{t-1}) = P(X_t = x_t | X_{t-1} = x_{t-1}, Y_{t-1} = y_{t-1})$,*
- *the conditional distribution of Y_t given the values of $\{Y_{t'}\}_{t' < t}$ and $\{X_{t'}\}_{t' \leq t}$ only depends on X_t and Y_{t-1}, \dots, Y_{t-s} and this conditional distribution has a probability density function (p.d.f.) $p_2(y_t | x_t, y_{t-s}^{t-1})$.*

Let us write $p(\cdot | x_{t-u}^{t-1}, y_{t-u}^{t-1})$ for the conditional p.d.f. of (X_t, Y_t) given $(X_{t-u}^{t-1} = x_{t-u}^{t-1}, Y_{t-u}^{t-1} = y_{t-u}^{t-1})$. Hypothesis 1 implies that for $u \geq s$

$$p(x_t, y_t | x_{t-u}^{t-1}, y_{t-u}^{t-1}) = p_1(x_t | x_{t-1}, y_{t-1}) p_2(y_t | x_t, y_{t-s}^{t-1}). \quad (1)$$

The various conditional independence assumptions are summarized by the directed graph below for $s = 1$.

$$\begin{array}{ccccccc}
\cdots & \rightarrow & X_{t-1} & \rightarrow & X_t & \rightarrow & X_{t+1} & \rightarrow & \cdots \\
& & \downarrow & \nearrow & \downarrow & \nearrow & \downarrow & & \\
\cdots & \rightarrow & Y_{t-1} & \rightarrow & Y_t & \rightarrow & Y_{t+1} & \rightarrow & \cdots
\end{array}$$

NHMS-AR models define a quite general family of models:

- If $p_1(x_t|x_{t-1}, y_{t-1})$ does not depend on y_{t-1} , we retrieve the usual MS-AR models which include the HMMs as a particular case ($s = 0$).
- If $M = 1$, $\{Y_t\}$ is an autoregressive process of order s .
- If $p_{1,\theta}(x_k|x_{k-1}, y_{k-s}^{k-1})$ does not depend on x_{k-1} and is parametrized using indicator functions, we obtain the Threshold AutoRegressive (TAR) models which is another important family of models with regime-switching in the literature (see e.g. [23]).

The following sections propose specific parametric models for p_1 (see Section 2.2) and p_2 when using Cartesian coordinates (see 2.3) or polar coordinates (see Sections 2.4 and 2.5).

2.2 Non-homogeneous Markov model for the weather type

As mentioned earlier, we introduce the latent process $\{X_t\}$ to describe the weather type which evolution may depend on previous wind direction. For example, we expect that the probability of switching from the cyclonic to the anticyclonic conditions generally is higher when the wind is blowing from the north than when it is blowing from the South. Such features can be modeled through the transition kernel p_1 . Hereafter we assume that

$$p_1(x_t|x_{t-1}, \phi_{t-1}) \propto q_{x_{t-1}, x_t} f_{VM}(\phi_{t-1}; \lambda_{x_{t-1}, x_t}, \psi_{x_{t-1}, x_t}), \quad (2)$$

where $f_{VM}(\cdot; \kappa, \phi)$ is the p.d.f. of the von Mises distribution, ϕ_{t-1} is the wind direction at time $t - 1$, $Q = (q_{x, x'})_{x, x' \in \{1, \dots, M\}}$ is a stochastic matrix and, for $x, x' \in \{1, \dots, M\}$, $\lambda_{x, x'} \geq 0$ and $\psi_{x, x'} \in \mathbb{T}$ are unknown parameters. The von Mises distribution is a natural distribution for circular variables (see [19]) which p.d.f. with respect to the Lebesgue measure on \mathbb{T} , is given by

$$\forall z \in \mathbb{T}, \quad f_\gamma(z) := f_{VM}(z; \kappa, \phi) = \frac{1}{2\pi I_0(\kappa)} \exp(\kappa \cos(z - \phi)) = \frac{1}{2\pi I_0(|\gamma|)} \left| e^{\gamma e^{-iz}} \right|, \quad (3)$$

with $\gamma := \kappa e^{i\phi}$ a complex parameter. I_0 denotes the modified Bessel function of order 0 defined as

$$I_0(\kappa) := \frac{1}{2\pi} \int_{\mathbb{T}} \exp(\kappa \cos(z)) dz.$$

$\phi \in \mathbb{T}$ corresponds to the circular mean of the distribution and $\kappa \geq 0$ describes the concentration of the distribution around ϕ : when $\kappa = 0$ we get the uniform distribution whereas the larger is κ the more concentrated around ϕ the distribution is. This distribution is denoted by $VM(\gamma)$ hereafter.

According to (2), the probability that the hidden Markov chain $\{X_t\}$ switches from x_{t-1} to x_t will increase when the wind direction Φ_{t-1} is close to ψ_{x_{t-1}, x_t} and λ_{x_{t-1}, x_t} models the directional

spreading in which this transition is likely to occur. When $\lambda_{x,x'} = 0$ for all $x, x' \in \{1, \dots, M\}$ we obtain again the homogeneous MS-AR models. Observe that (2) can be rewritten

$$p_1(x_t|x_{t-1}, \phi_{t-1}) = \frac{q_{x_{t-1}, x_t} \left| \exp \left(\tilde{\lambda}_{x_{t-1}, x_t} e^{-i\phi_{t-1}} \right) \right|}{\sum_{x'=1}^M q_{x_{t-1}, x'} \left| \exp \left(\tilde{\lambda}_{x_{t-1}, x'} e^{-i\phi_{t-1}} \right) \right|}, \quad (4)$$

with $\tilde{\lambda}_{x,x'} \in \mathbb{C}$ (by taking $\tilde{\lambda}_{x,x'} = \lambda_{x,x'} e^{i\psi_{x,x'}}$). With this expression, it can be easily seen that replacing $(\lambda_{x,x'})_{x,x'}$ by $(\tilde{\lambda}_{x,x'} - a_x)_{x,x'}$ (for any choice of $(a_x)_x$) does not change p_1 and thus that identifiability constraints are needed.

In order to reduce the number of unknown parameters we add the following constraints for the non-homogeneous models developed in the sequel

$$\tilde{\lambda}_{x,x'} = \tilde{\lambda}_{x'} \quad (5)$$

for all $x, x' \in \{1, \dots, M\}$ such that $x \neq x'$ with the identifiability constraint

$$\sum_{x'=1}^M \tilde{\lambda}_{x'} = 0 \quad (6)$$

We have also fitted the model without the constraint (5) and found that the likelihood of these models is similar to the one of the models with the constraint (5) whereas they have a significantly larger number of parameters. Even when assuming (5), we found that the parameter $(\lambda_{x'})$ is sometimes hard to fit in practice and that fixing its values to e.g. the concentration parameter of the von-Mises distribution fitted to the time series of wind direction leads to satisfactory models. The results obtained with these alternative strategies are not further discussed below.

2.3 Modeling the Cartesian coordinates conditionally to the weather type

In this section we propose a model for the bivariate process $\{Y_t\} = \{u_t, v_t\}$ conditionally to the weather type $\{X_t\}$. This process has values in \mathbb{R}^2 and the most classical autoregressive model for such process is the linear Gaussian vector autoregressive (VAR) model of order s . With this model, if $X_t = x_t$ then

$$Y_t = A_0^{(x_t)} + A_1^{(x_t)} Y_{t-1} + \dots + A_s^{(x_t)} Y_{t-s} + \left(\Sigma^{(x_t)} \right)^{\frac{1}{2}} \epsilon_t \quad (7)$$

where $A_0^{(x)} \in \mathbb{R}^2$, $A_l^{(x)} \in \mathbb{R}^{2 \times 2}$ for $l \in \{1, \dots, s\}$ and $x \in \{1, \dots, M\}$, $\Sigma^{(x)} \in \mathbb{R}^{2 \times 2}$ are symmetric positive matrices for $x \in \{1, \dots, M\}$ and $\{\epsilon_t\}$ is a bivariate white noise sequence.

VAR models have been proposed for wind fields in a space-time context in [2, 25, 13]. On our particular dataset, we found that this model was not appropriate to reproduce the 'hole' around the origin which can be seen on the joint distribution on Figure 1. This hole corresponds to a low probability of observing low wind speed. We can get around this issue by applying a power transformation as follows

$$\begin{cases} \tilde{u}_t &= U_t^\alpha \cos(\Phi_t) \\ \tilde{v}_t &= U_t^\alpha \sin(\Phi_t) \end{cases}$$

and fit the MS-AR model to $\{\tilde{u}_t, \tilde{v}_t\}$ instead of $\{u_t, v_t\}$. The value $\alpha = 1.5$ was chosen experimentally to remove the 'hole' close to the origin in the original distribution. The model with homogeneous hidden Markov chain is denoted **HMS-AR**_(u,v) and the non-homogeneous model, where p_1 is given by (2), is denoted **NHMS-AR**_(u,v).

2.4 Modeling the wind direction conditionally to the weather type

In this section we propose a model for the circular process $\{Y_t\} = \{\Phi_t\}$. The inclusion of $\{U_t\}$ in this model is discussed in the next section. Several autoregressive models have been proposed in the literature for modeling directional time series (see [5], [11],[15], [17]). We have chosen to focus on the von Mises process initially introduced in [5] and assume that the conditional distribution of Y_t given $(X_t = x_t, Y_{t-s}^{t-1} = y_{t-s}^{t-1})$ is $VM\left(\gamma_0^{(x_t)} + \sum_{\ell=1}^s \gamma_\ell^{(x_t)} e^{iy_{t-\ell}}\right)$ with $\gamma_\ell^{(x)} = \kappa_\ell^{(x)} e^{i\phi_\ell^{(x)}} \in \mathbb{C}$ for $x \in \{1, \dots, M\}$ and $\ell \in \{0, \dots, s\}$. This can be rewritten

$$\begin{aligned} p_2(y_t|x_t, y_{t-s}^{t-1}) &= \frac{1}{b(x_t, y_{t-s}^{t-1})} \exp\left(\kappa_0^{(x_t)} \cos(y_t - \phi_0^{(x_t)}) + \sum_{\ell=1}^s \kappa_\ell^{(x_t)} \cos(y_t - y_{t-\ell} - \phi_\ell^{(x)})\right) \\ &= \frac{1}{b(x_t, y_{t-s}^{t-1})} \left| \exp\left([\gamma_0^{(x_t)} + \sum_{\ell=1}^s \gamma_\ell^{(x_t)} e^{iy_{t-\ell}}] e^{-iy_t}\right) \right| \end{aligned} \quad (8)$$

with

$$\begin{aligned} b(x_t, y_{t-s}^{t-1}) &:= \int_{\mathbb{T}} \exp\left(\kappa_0^{(x_t)} \cos(y - \phi_0^{(x_t)}) + \sum_{\ell=1}^s \kappa_\ell^{(x_t)} \cos(y - y_{t-\ell})\right) dy \\ &= I_0\left(\left|\gamma_0^{(x_t)} + \sum_{\ell=1}^s \gamma_\ell^{(x_t)} e^{iy_{t-\ell}}\right|\right). \end{aligned}$$

In [5], it was assumed that $\gamma_\ell^{(x)} \in \mathbb{R}$ for $\ell \in \{1, \dots, s\}$. We have chosen to extend it to a model with complex parameters in order to be able to reproduce the prevailing rotation of the wind direction in the clockwise direction (see Section 4).

The model with homogeneous hidden Markov chain is denoted **HMS-EVM** and the non-homogeneous model, where p_1 is given by (2), is denoted **NHMS-EVM**.

2.5 A conditional model for the wind speed given the wind direction

In [1] it was proposed to model the wind speed $\{U_t\}$ using a homogeneous MS-AR model with $M = 3$ regimes and Gaussian linear AR models (see (7)) of order $s = 2$. Figure 2 shows typical examples of wind speed and wind direction time series together with the regimes identified by the fitted MS-AR models. These regimes basically correspond to periods with different temporal variability and there seems to be no simple relation between the regimes identified on the wind speed and the wind direction. In this context, it does not seem relevant to use the same weather type for the two time series. We thus propose to introduce a different weather type $X_t^{(U)}$ for the wind speed and $X_t^{(\Phi)}$ for the wind direction.

In order to explore the link between $X_t^{(U)}$ and Φ , we have computed the most likely values of $X_t^{(U)}$ given the observed time series of wind speed $\{U_t\}$ and produced rose plots of the wind direction in the different weather types which were identified for the wind speed. We got plots very similar to the ones shown on Figure 3 (right panel). The first regime, which corresponds to periods with low temporal variability for the wind speed (anticyclonic conditions), can occur in any wind direction whereas the more variable regimes 2 and 3 are mainly associated to south-westerlies (cyclonic conditions). Stated otherwise the wind direction provides information on the

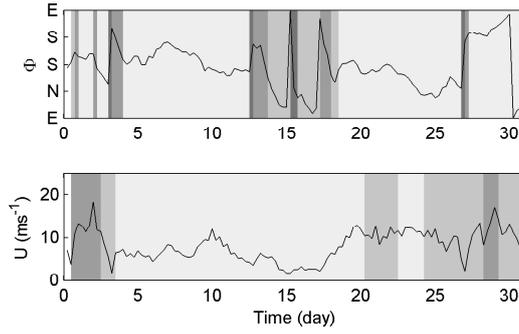
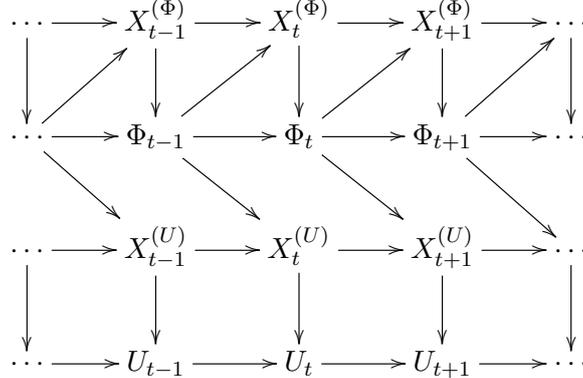


Figure 2: Example of time series of wind direction (top plot) and wind speed (bottom plot). The colors indicate the most likely regimes for the fitted **NHMS-EVM** model with 4 regimes (top plot) and Gaussian homogeneous MS-AR model for wind speed with 3 regimes (bottom plot). The regimes have been ordered according to the time variability (the darker the more variability).

synoptic weather conditions which control the intensity and the variability of the wind speed. It suggests the use of a non-homogeneous MS-AR model for the wind speed where the transition probabilities depend on the wind direction. Hereafter **NHMS-AR** $_{(U,\Phi)}$ denotes the model for $\{U_t, \Phi_t\}$ such that

- $\{\Phi_t\}$ is modeled by the **NHMS-EVM** model for $\{x_t^{(\Phi)}, \Phi_t\}$ with $M = 4$ and $s = 2$,
- $\{U_t\}$ is modeled conditionally to $\{\Phi_t\}$ by a NHMS-AR model with p_1 given by (2) and p_2 by a linear Gaussian AR model (7).

The structure of the model, with two layers of hidden variables, one for the wind speed and one for the wind direction, is shown on the directed graph below.



3 Parameter estimation

3.1 Numerical computation of the MLE

The parameter vector of NH-MSAR models is composed of the parameters θ_Q of the transition probabilities $p_1(x_t|x_{t-1}, y_{t-1})$, the parameters $\theta^{(x)}$ of the transition kernel $p_2(y_t|x, y_{t-s}^{t-1})$ for each regime $x \in \{1, \dots, M\}$.

They are estimated by maximizing the likelihood function using a generalized EM algorithm. This algorithm was initially introduced in [4] for HMMs and then generalized to models with latent variables in [9]. This recursive algorithm computes successive approximations $\hat{\theta}_i$ of the maximum likelihood estimate (MLE) $\hat{\theta}$ by cycling through the following steps.

E-step: Compute $Q(\theta|\hat{\theta}_i) = E_{\hat{\theta}_i}(\log(p_\theta(X_1^T, Y_1^T))|y_{-s+1}^T)$ as a function of θ .

M-step: Determine the updated parameter estimate $\hat{\theta}_{i+1} = \arg \max_{\theta} Q(\theta|\hat{\theta}_i)$.

The conditional probabilities involved in the computation of $Q(\theta|\hat{\theta}_n)$ are computed using the so-called forward-backward recursions (see e.g. [7] and references therein). The particular implementation of these recursions for homogeneous MS-AR models is discussed in [14] and [16] discusses it for non-homogeneous HMMs. It can be easily generalized to the models considered in this paper. The M-step requires numerical optimization leading to the so-called Generalized EM (GEM) algorithm. In order to get an efficient EM algorithm, it is important to implement carefully the numerical optimization procedure. In practice, the function $Q(\cdot|\hat{\theta}_i)$ which has to be maximized in the M-step can be written as the sum of $M + 1$ functions as follows

$$Q(\theta|\hat{\theta}_i) = Q_X(\theta_Q|\hat{\theta}_i) + \sum_{x=1}^M Q_Y(\theta^{(x)}|\hat{\theta}_i).$$

This leads to solving $M + 1$ separate optimization problems on spaces with reduced dimension which is far more efficient than maximizing directly $Q(\cdot|\hat{\theta}_i)$ over all the parameters. Note that analytical expressions are available for $\arg \max_{\theta^{(x)}} Q_Y(\theta^{(x)}|\hat{\theta}_i)$ when Gaussian linear AR models (7) are used and for $\arg \max_{\theta_Q} Q_X(\theta_Q|\hat{\theta}_i)$ when homogeneous MS-AR models are considered.

In order to avoid convergence to non-interesting maxima and save computational time, a proper initialization of this algorithm with realistic parameter values $\hat{\theta}_0$ is needed. In practice, we have used the nested nature of the models. We have first fitted homogeneous models and then use the estimated parameters as a starting point for the corresponding non-homogeneous models. In the same spirit, the results obtained for the model of order s have been used to initialize the EM for the models of order $s + 1$.

3.2 Ergodicity of the models and asymptotic properties of the MLE

Conditions on the parameters which warrant the existence of a unique stationary ergodic solution with finite moments of order 2 are given in [3] for the models with linear Gaussian AR models (7). These are desirable properties for the fitted models since we expect that the wind process satisfies such properties and imply that the MLE are consistent (see [3]). These conditions apply to the models with linear Gaussian AR discussed in Section 4 and thus the MLE are consistent for these models. In this section we show that similar results hold true for the model with von Mises distribution introduced in Section 2.4 and prove more precisely ψ -irreducibility, aperiodicity, Harris-recurrence, identifiability and consistency for the NHMS-EVM model.

In the rest of this section, we assume Hypothesis 1 with p_1 and p_2 given by (4) and (8) respectively. Let Θ' be the set of parameters $\theta = (q_{x,x'}, \tilde{\lambda}_{x,x'}, \gamma_\ell^{(x)})_{\ell,x,x'}$ such that $\gamma_\ell^{(x)} \in \mathbb{C}$, $q_{x,x'} > 0$ such that $\sum_{x'} q_{x,x'} = 1$ and $\tilde{\lambda}_{x,x'} \in \mathbb{C}$ satisfying (6). Let us now state our main result.

Theorem 2 (Consistency for NHMS-EVM). *Assume that Θ is a compact subset of Θ' and that*

the coordinates of the true parameter θ^* satisfy

$$x \neq x' \Rightarrow (\gamma_{0,*}^{(x)}, \dots, \gamma_{s,*}^{(x)}) \neq (\gamma_{0,*}^{(x')}, \dots, \gamma_{s,*}^{(x')}). \quad (9)$$

Then, for every $x_0 \in \{1, \dots, M\}$ and any initial measure ν on $\{1, \dots, M\} \times \mathbb{T}$, on a set of probability one, the limit values $\theta = (\gamma, Q, \tilde{\lambda})$ of the sequence of MLE $(\hat{\theta}_{n,x_0})_n$ are equal to $\theta^* = (\gamma_*, Q_*, \tilde{\lambda}_*)$ up to a permutation of indices, i.e. for any such limit value θ , there exists a permutation σ of $\{1, \dots, M\}$ such that, for every $x, x' \in \{1, \dots, M\}$, for every $j = 0, \dots, s$, the following relations hold true

$$\gamma_j^{(x)} = \gamma_{j,*}^{(\sigma(x))}, \quad q_{x,x'} = q_{\sigma(x),\sigma(x'),*} \quad \text{and} \quad \tilde{\lambda}_{x,x'} = \tilde{\lambda}_{\sigma(x),\sigma(x'),*}.$$

One can notice that (9) just means that there is no couple of regimes (x, x') with $x \neq x'$ in which the behavior of the process $\{Y_t\}$ is the same.

The proof of Theorem 2 is based on two ingredients: a general consistency result established in [3][Thm 2] and the proof of the identifiability up to a permutation of indices (see Proposition 4). Let us first check that the conditions of [3][Thm 2] apply for the NHMS-EVM model.

Since for every $(\theta, x, y) \in \Theta \times \{1, \dots, M\} \times \mathbb{T}$, $q_\theta(x, y | \cdot, \cdot)$ is continuous on the compact set $\{1, \dots, M\} \times \mathbb{T}^s$, we have

$$\alpha = \int_{E \times K} \gamma(x, y) d\mu_0(x, y) > 0, \quad \text{with} \quad \gamma(x, y) := \inf_{x', y_{-s}^{-1}} q_\theta(x, y | x', y_{-s}^{-1}).$$

Now we consider the probability density function (w.r.t. μ_0) β given by

$$\beta(x, y) := \frac{\gamma(x, y)}{\alpha}.$$

For every $x_0, x_{-1} \in E$ and every y_{-s}^0 , we have

$$q_\theta(x_0, y_0 | x_{-1}, y_{-s}^{-1}) \geq \alpha \beta(x_0, y_0).$$

This implies the ψ -irreducibility, the strong aperiodicity (the ν_1 -small set being the whole space), the Harris recurrence (since we can decompose the whole set in a union of uniformly accessible sets from the whole set) and the positiveness (the invariant measure being unique and finite). In particular, this gives Assumption (5) of [3][Thm 2].

Moreover, since $p_{1,\theta}(x_1 | x_0, y_0)$ and $p_{2,\theta}(y_0 | x_0, y_{-1})$ are continuous in (θ, x_1, x_0, y_0) and in $(\theta, x_0, y_0, y_{-1})$ (respectively), all the other assumptions of [3][Thm 2] are satisfied for any compact subset of Θ' . Hence, we have

Corollary 3. *Assume that Θ is a compact subset of Θ' . Then, for all $\theta \in \Theta$, there exists a unique invariant probability and, for every $x_0 \in E$ and every initial probability ν , the limit values of $(\hat{\theta}_{n,x_0})_n$ are $\bar{\mathbb{P}}_{\theta^*}$ -almost surely contained in $\{\theta \in \Theta : \bar{\mathbb{P}}_\theta = \bar{\mathbb{P}}_{\theta^*}\}$.*

Now, Theorem 2 will follow from the following result giving the identifiability of the parameter (up to permutation of indices) .

Proposition 4 (Identifiability). *Let θ_1 and θ_2 belong to Θ' :*

$$\theta_i = \left((\gamma_{j,(i)}^{(x)})_{j,x}, (q_{x,x',(i)})_{x,x'}, (\tilde{\lambda}_{x,x',(i)})_{x,x'} \right).$$

Assume that the parameters $(\gamma_{j,(1)}^{(x)})_{j,x}$ which model the evolution of the wind direction in the different regimes through (8) for θ_1 are such that

$$x \neq x' \Rightarrow (\gamma_{0,(1)}^{(x)}, \dots, \gamma_{s,(1)}^{(x)}) \neq (\gamma_{0,(1)}^{(x')}, \dots, \gamma_{s,(1)}^{(x')}). \quad (10)$$

Then $\bar{\mathbb{P}}_{\theta_1}^Y = \bar{\mathbb{P}}_{\theta_2}^Y$ if and only θ_1 is equal to θ_2 up to a permutation of indices.

Proof of Proposition 4. We write Leb for the Lebesgue measure on \mathbb{T} . Assume that $\bar{\mathbb{P}}_{\theta_1}^Y = \bar{\mathbb{P}}_{\theta_2}^Y$. In particular, we have

$$\bar{p}_{\theta_1}(Y_t = y_t | Y_{t-s}^{t-1} = y_{t-s}^{t-1}) = \bar{p}_{\theta_2}(Y_t = y_t | Y_{t-s}^{t-1} = y_{t-s}^{t-1}), \text{ for } \bar{\mathbb{P}}_{\theta_1}^{Y_{t-s}^t} - a.e. y_{t-s}^t$$

and thus

$$\sum_{x=1}^M \bar{\mathbb{P}}_{\theta_1}(X_t = x | y_{t-s}^{t-1}) p_{2,\theta_1}(y_t | x, y_{t-s}^{t-1}) = \sum_{x=1}^M \bar{\mathbb{P}}_{\theta_2}(X_t = x | y_{t-s}^{t-1}) p_{2,\theta_2}(y_t | x, y_{t-s}^{t-1}), \text{ for } \bar{\mathbb{P}}_{\theta_1}^{Y_{t-s}^t} - a.e. y_{t-s}^t.$$

Since $\bar{p}_{\theta_1}(y_{t-s}^t) > 0$ (since the invariant density h_1 satisfies $h_1 > 0$ since $\alpha > 0$ and the transition density p satisfies $p > 0$ by construction) and since (4) holds true, we deduce that, for $\text{Leb}^{\otimes(s+1)}$ -a.e. y_{t-s}^t , we have

$$\sum_{x=1}^M \bar{\mathbb{P}}_{\theta_1}(X_t = x | y_{t-s}^{t-1}) f_{\gamma_{0,(1)}^{(x)} + \sum_{\ell=1}^s \gamma_{\ell,(1)}^{(x)}}(y_t) = \sum_{x=1}^M \bar{\mathbb{P}}_{\theta_2}(X_t = x | y_{t-s}^{t-1}) f_{\gamma_{0,(2)}^{(x)} + \sum_{\ell=1}^s \gamma_{\ell,(2)}^{(x)}}(y_t)$$

with f_γ defined by (3). Due to [12], finite mixtures of von Mises distributions are identifiable. Hence if

$$\sum_{x=1}^M \pi_1^{(x)} f_{\gamma_1^{(x)}}(y) = \sum_{x=1}^M \pi_2^{(x)} f_{\gamma_2^{(x)}}(y) \text{ for Leb } -a.e. y$$

with $\gamma_1^{(x)} \neq \gamma_1^{(x')}$ for $x \neq x'$ and $\pi_1^{(x)} > 0$ for $x \in \{1, \dots, M\}$ then there exists a permutation $\tau : \{1, \dots, M\} \rightarrow \{1, \dots, M\}$ such that $\gamma_1^{(x)} = \gamma_2^{(\tau(x))}$ and $\pi_1^{(x)} = \pi_2^{(\tau(x))}$.

Recall that we have assumed that $\theta_{Y,1}^{(x)} \neq \theta_{Y,1}^{(x')}$ if $x \neq x'$, which implies that

$$\text{for } \text{Leb}^{\otimes s} - a.e. y_{t-s}^{t-1}, \quad \gamma_{0,1}^{(x)} + \sum_{\ell=1}^s \gamma_{\ell,1}^{(x)} e^{iy_{t-\ell}} \neq \gamma_{0,1}^{(x')} + \sum_{\ell=1}^s \gamma_{\ell,1}^{(x')} e^{iy_{t-\ell}}.$$

Therefore, since for every $x \in \{1, \dots, M\}$ and for $\text{Leb}^{\otimes s}$ -almost every y_{t-s}^{t-1} , $\bar{\mathbb{P}}_{\theta_1}(X_t = x | y_{t-s}^{t-1}) > 0$ (since $h_{\theta_1} > 0$), for $\text{Leb}^{\otimes s}$ -almost every y_{t-s}^{t-1} there exists a permutation $\sigma_{y_{t-s}^{t-1}}$ of $\{1, \dots, M\}$ such that,

$$\forall x \in \{1, \dots, M\}, \quad \gamma_{0,(1)}^{(x)} + \sum_{\ell=1}^s \gamma_{\ell,(1)}^{(x)} e^{iy_{t-\ell}} = \gamma_{0,(2)}^{(\sigma_{y_{t-s}^{t-1}}(x))} + \sum_{\ell=1}^s \gamma_{\ell,(2)}^{(\sigma_{y_{t-s}^{t-1}}(x))} e^{iy_{t-\ell}}.$$

Since the set of permutations of $\{1, \dots, M\}$ is finite, there exists a positive Lebesgue measure subset of \mathbb{T}^s on which the permutation is the same permutation σ . From this, we deduce that

$$\forall x \in \{1, \dots, M\}, \quad \forall j \in \{0, \dots, s\}, \quad \gamma_{j,(1)}^{(x)} = \gamma_{j,(2)}^{(\sigma(x))}$$

and that, for Lebesgue almost every y_{t-s}^{t+1} , the following holds true

$$\forall x \in \{1, \dots, M\}, \quad \bar{\mathbb{P}}_{\theta_1}(X_t = x | y_{t-s}^{t-1}) = \bar{\mathbb{P}}_{\theta_2}(X_t = \sigma(x) | y_{t-s}^{t-1}).$$

Let us now discuss the identifiability of the other components of θ_1 and θ_2 . If $\bar{\mathbb{P}}_{\theta_1}^Y = \bar{\mathbb{P}}_{\theta_2}^Y$ then

$$\bar{p}_{\theta_1}(Y_t = y_t, Y_{t+1} = y_{t+1} | Y_{t-s}^{t-1} = y_{t-s}^{t-1}) = \bar{p}_{\theta_2}(Y_t = y_t, Y_{t+1} = y_{t+1} | Y_{t-s}^{t-1} = y_{t-s}^{t-1}) \bar{\mathbb{P}}_{\theta_1}^Y \text{ a.s.}$$

and thus, for Lebesgue almost every y_{t-s}^{t+1} , we have

$$\begin{aligned} & \sum_{x, x'=1}^M \bar{\mathbb{P}}_{\theta_1}(X_t = x | y_{t-s}^{t-1}) p_{1, \theta_1}(x' | x, y_t) f_{\gamma_{0, (1)} + \sum_{\ell=1}^s \gamma_{\ell, (1)}^{(x)}} e^{iy_t - \ell}(y_t) f_{\gamma_{0, (1)} + \sum_{\ell=1}^s \gamma_{\ell, (1)}^{(x')}} e^{iy_t - \ell + 1}(y_{t+1}) \\ &= \sum_{x, x'=1}^M \bar{\mathbb{P}}_{\theta_2}(X_t = x | y_{t-s}^{t-1}) p_{1, \theta_2}(x' | x, y_t) f_{\gamma_{0, (2)} + \sum_{\ell=1}^s \gamma_{\ell, (2)}^{(x)}} e^{iy_t - \ell}(y_t) f_{\gamma_{0, (2)} + \sum_{\ell=1}^s \gamma_{\ell, (2)}^{(x')}} e^{iy_t - \ell + 1}(y_{t+1}). \end{aligned}$$

Using again the identifiability of von Mises distribution, we obtain

$$\forall x, x', \quad p_{1, \theta_1}(x' | x, y_t) = p_{1, \theta_2}(\sigma(x') | \sigma(x), y_t) \text{ for Leb-a.e. } y_t.$$

Now, due to the special form of $p_{1, \theta}$ specified in (4), we get

$$\forall x, x', \text{ Leb-a.e. } y_t, \quad \frac{q_{x, x', (1)} \left| \exp \left(\tilde{\lambda}_{x, x', (1)} e^{-iy_t} \right) \right|}{\sum_{x''=1}^M q_{x, x'', (1)} \left| \exp \left(\tilde{\lambda}_{x, x'', (1)} e^{-iy_t} \right) \right|} = \frac{q_{\sigma(x), \sigma(x'), (2)} \left| \exp \left(\tilde{\lambda}_{\sigma(x), \sigma(x'), (2)} e^{-iy_t} \right) \right|}{\sum_{x''=1}^M q_{\sigma(x), x'', (2)} \left| \exp \left(\tilde{\lambda}_{\sigma(x), x'', (2)} e^{-iy_t} \right) \right|}. \quad (11)$$

Let $x \in \{1, \dots, M\}$ be fixed. Applying (11) a first time with $x' = x$ and a second time with any x' , we get

$$\forall x', \text{ for Leb-a.e. } y_t, \quad \frac{q_{x, x', (1)} \left| \exp \left(\tilde{\lambda}_{x, x', (1)} e^{-iy_t} \right) \right|}{q_{x, x, (1)} \left| \exp \left(\tilde{\lambda}_{x, x, (1)} e^{-iy_t} \right) \right|} = \frac{q_{\sigma(x), \sigma(x'), (2)} \left| \exp \left(\tilde{\lambda}_{\sigma(x), \sigma(x'), (2)} e^{-iy_t} \right) \right|}{q_{\sigma(x), \sigma(x), (2)} \left| \exp \left(\tilde{\lambda}_{\sigma(x), \sigma(x), (2)} e^{-iy_t} \right) \right|}$$

and so

$$\forall x', \quad \frac{q_{x, x', (1)}}{q_{x, x, (1)}} = \frac{q_{\sigma(x), \sigma(x'), (2)}}{q_{\sigma(x), \sigma(x), (2)}} \quad (12)$$

and

$$\forall x', \quad \tilde{\lambda}_{x, x', (1)} - \tilde{\lambda}_{x, x, (1)} = \tilde{\lambda}_{\sigma(x), \sigma(x'), (2)} - \tilde{\lambda}_{\sigma(x), \sigma(x), (2)}. \quad (13)$$

Now, since $\sum_{x'} q_{x, x', (1)} = 1 = \sum_{x'} q_{\sigma(x), \sigma(x'), (2)}$, due to (12), it comes $q_{x, x, (1)} = q_{\sigma(x), \sigma(x), (2)}$ and so

$$\forall x' \in E, \quad q_{x, x', (1)} = q_{\sigma(x), \sigma(x'), (2)}.$$

Since θ_1 and θ_2 are in Θ' , $\sum_{x'} \tilde{\lambda}_{x, x', (1)} = 0 = \sum_{x'} \tilde{\lambda}_{\sigma(x), \sigma(x'), (2)}$, and due to (13), we get $\tilde{\lambda}_{x, x, (1)} = \tilde{\lambda}_{\sigma(x), \sigma(x), (2)}$ and, applying again (13), we conclude that

$$\forall x' \in E, \quad \tilde{\lambda}_{x, x', (1)} = \tilde{\lambda}_{\sigma(x), \sigma(x'), (2)}.$$

□

4 Numerical results

4.1 Data

In this paper, we focus on a wind time series extracted from 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>. We consider the surface wind data for the point with geographical coordinates (47.5° N, 5° W) which is located off the Brittany coast (northwest of France). We have performed a comparison with in-situ data which indicates that this reanalysis data provides an accurate description of the wind condition observed at this location with the advantage of being easy to use in a statistical study (long time series with no missing data). The resulting time series is non-stationary since it exhibits an important seasonal component but also diurnal and inter-annual components. A classical approach for treating seasonality in meteorological time series consists in blocking the data, typically by period ranging from a month to a trimester depending on the amount of data available, and in fitting a separate model for each period in the year. This approach has been used in this paper and we have chosen to focus on the months of January. It leads to 44 time series of length 124 (31 days with 4 observations each day), each time series describing the wind conditions during the month of January for a particular year. In the sequel, we assume that these time series are independent realizations of a stationary process. It seems realistic according to the results given in [1] for the wind speed at the same location since the diurnal components can be neglected during the winter season. Following [1], another approach would consist in letting some of the coefficients of the models introduced hereafter to evolve in time with periodic functions for the diurnal and seasonal components and eventually a trend.

4.2 Model selection

Before analyzing any numerical result, one has to discuss the choice of the number of regimes and of the order of the AR models. In practice we found that the BIC criterion generally permits to identify parsimonious models which fit well the data. It is defined as

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

and L is the likelihood of the data, k is the number of parameters and N is the number of observations. In order to make the final selection among the best models identified by BIC, we have compared their abilities to generate realistic wind time series since this is the main motivation for this work. For this, a large number of realizations of the various models under competition have been simulated and various statistics of these synthetic sequences have been compared with the ones of the original data.

The models were fitted with a number of regimes M varying from 1 to 6 at the most and the BIC values suggest selecting models with $M = 3$ or $M = 4$ regimes (see Tables 1 and 2). For the wind direction the model with $M = 4$ regimes tends to better reproduce the marginal distribution of the process compared to the models with $M = 3$ regimes and we thus have chosen to select this model. The **NHMS-EVM** model with $M = 4$ and $s = 2$, which is used in the **NHMS-AR**_(U,Φ) model, has 43 parameters. For the Cartesian coordinates $\{u_t, v_t\}$ and for the wind intensity $\{U_t\}$ the models with $M = 3$ and $M = 4$ regimes lead to similar results and we have thus chosen to keep the simplest model with $M = 3$.

We also varied the order s of the autoregressive models from $s = 0$ (y_t is independent of y_0^{t-1} given x_t) to $s = 5$ and the BIC values are generally decreasing with s suggesting that a model of order $s \geq 5$ may be needed. Notice however that there is generally a big improvement in the BIC values when s increases from 0 to 1 and from 1 to 2 whereas the difference is much smaller when comparing the models of order $s = 2$ and $s \geq 3$ (not shown). We will focus on models of order $s = 2$ in the sequel. We believe that models of reduced order are more realistic from a

physical point of view and we get similar simulation results with $s = 2$ compared to the models with $s \geq 3$.

The BIC of $\{u_t, v_t\}$ models are generally smaller than the ones of $\{U_t, \Phi_t\}$ models except for $s = 0$. It may be due to the higher number of parameters involved in the **NHMS-AR** $_{(U,\Phi)}$ model (66 parameters when $M = 3$), which has two layers of hidden variables whereas the **NHMS-AR** $_{(u,v)}$ model has one common weather type for $\{u_t\}$ and $\{v_t\}$ and only 44 parameters when $M = 3$. Note however that BIC does not permit to make a clear distinction between both parameterizations (polar or Cartesian) since the differences in BIC values are relatively small.

	M	1	2	3	4	5	6	k
Model	s	BIC						
HMS-EVM	1	7778	6326	6334	6307	6277	6385	$M(M-1)+4M$
NHMS-EVM	1	7778	6266	6171	6141	6158	6372	$M(M+1)-1+4M$
HMS-EVM	2	7568	5952	5979	5963	6051	6075	$M(M-1)+6M$
NHMS-EVM	2	7568	5882	5872	5882	5968	6075	$M(M+1)-1+6M$

Table 1: BIC values for the various fitted wind direction models. The last column gives the number of parameters. The bold value corresponds to the selected model.

	M	1	2	3	4	5	k
Model	s	BIC					
HMS-AR $_{(u,v)}$	0	48583	44616	42338	40903	40025	$M(M-1)+5M$
NHMS-AR $_{(u,v)}$	0	-	43679	41212	39878		$M(M+1)-1+5M$
NHMS-AR $_{(U,\Phi)}$	0	-	32553	31180	30381	30000	$43+M(M+1)-1+2M$
HMS-AR $_{(u,v)}$	1	31979	28134	27561	27219	27079	$(M(M-1)+9M)$
NHMS-AR $_{(u,v)}$	1	-	27687	27110	26855	26755	$M(M+1)-1+9M$
NHMS-AR $_{(U,\Phi)}$	1	-	28833	28543	28446	28380	$43+M(M+1)-1+3M$
HMS-AR $_{(u,v)}$	2	30619	26753	26162	25950	25947	$(M(M-1)+13M)$
NHMS-AR $_{(u,v)}$	2	-	26275	25681	25598	25607	$M(M+1)-1+13M$
NHMS-AR $_{(U,\Phi)}$	2	-	28458	28266	28163	28196	$43+M(M+1)-1+4M$

Table 2: BIC values for the various bivariate models. The last column gives the number of parameters. The bold values correspond to the selected models.

4.3 Regimes can be interpreted as weather types

An important benefit of using weather type models for meteorological variables is that they generally lead to interpretable models. This is illustrated in this section on **NHMS-AR** $_{(u,v)}$ and **NHMS-AR** $_{(U,\Phi)}$ models. In order to compare the regimes of these two models, they have been ordered according to the variance of the innovation of the autoregressive processes $\Sigma^{(s)}$. Figure 3 shows that the distributions of the wind direction in the different regimes are broadly similar for both models. The first regime corresponds mainly to anticyclonic conditions with easterly wind and a slow varying intensity (the variance of the innovation of the AR model is lower than in the other regimes and the first AR coefficient is larger). This regime is also the most likely (probability of occurrence of about 46%). The two other regimes correspond to cyclonic conditions with westerly wind and higher temporal variability in the intensity. These two regimes are discriminated mainly by the temporal variability, which is higher in the third

regime, and the wind direction with the second regime corresponding mainly to south-westerlies and the third regime corresponding mainly to north-westerlies (see Figure 3).

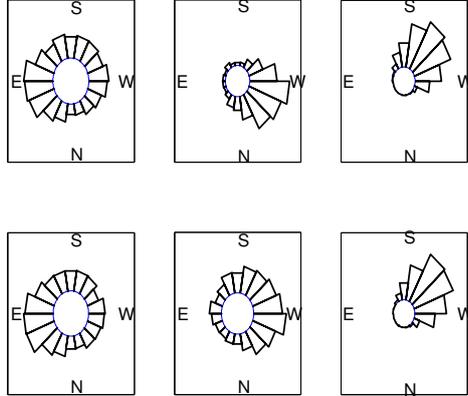


Figure 3: Rose plot of the wind direction in the three regimes identified on wind speed by the $\mathbf{NHMS-AR}_{(U,\Phi)}$ model (top) and the $\mathbf{NHMS-AR}_{(u,v)}$ model (bottom) .

In order to assess the physical consistency of these three local regimes, we confronted them to the large-scale regimes commonly used in climate studies. More precisely, we considered the four wintertime regimes of [8], obtained over the North-Atlantic/European sector (90W-30E / 20-80N) by a kmeans-clustering of 3607 daily maps of 500 mb geopotential anomalies (days of December, January and February 1981–2010). The classification of each day of January 1979–2001 into these four types was used to compute the conditional probabilities given in Table 3. It shows a clear link between the regimes identified by the $\mathbf{NHMS-AR}_{(u,v)}$ model and the large-scale regimes. For example the first regime of the $\mathbf{NHMS-AR}_{(u,v)}$ model, with low temporal variability, is more likely to occur when an anticyclone, generally located over Scandinavia, blocks the westerly flow (large scale regime denoted BL). At the opposite the more variable third regime of the $\mathbf{NHMS-AR}_{(u,v)}$ model is generally associated with the large scale regime NAO+ (positive phase of the North Atlantic Oscillation). The regimes AR (Atlantic Ridge) and NAO- (negative phases of the North Atlantic Oscillation) have intermediate temporal variability with AR being more stable than NAO-. These results are coherent with the climatology of the area.

Regime	BL	AR	NAO-	NAO+	Total
1	0.1832	0.0793	0.0621	0.0548	0.3794
2	0.0897	0.0919	0.0996	0.1728	0.4540
3	0.0080	0.0166	0.0316	0.1103	0.1666
Total	0.2810	0.1878	0.1933	0.3380	1

Table 3: Joint probability of occurrence of the three regimes identified by the the $\mathbf{NHMS-AR}_{(u,v)}$ model (lines) and the large-scale regimes provided by J. Cattiaux (see [8]) in columns.

Both $\mathbf{NHMS-AR}_{(u,v)}$ and $\mathbf{NHMS-AR}_{(U,\Phi)}$ models have similar transition probabilities (see Figure 4) with a more pronounced dependence on the wind direction for the $\mathbf{NHMS-AR}_{(U,\Phi)}$ model. The more persistent regime is clearly the first one (mean duration of about 3.4 days) with a high probability of staying in this regime in any wind direction. The probability of switching directly from regime 1 to regime 3 is very small and the Markov chain will generally

transit through the regime 2. Transitions from regime 1 to regime 2 are more likely when the wind is blowing from the west and transitions from regime 2 to regime 3 generally occur when the wind is from south. Regime 3 is persistent only when the wind is from south-west. If the wind blows from other directions, the weather type will quickly switch to regime 1 or 2.

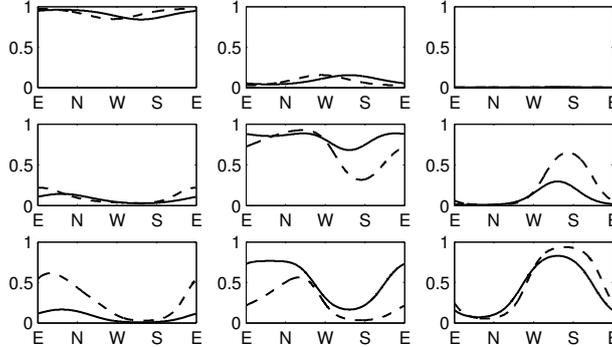


Figure 4: Matrix of non-homogeneous transitions of $\mathbf{NHMS-AR}_{(u,v)}$ model (plain line) and $\mathbf{NHMS-AR}_{(U,\Phi)}$ model (dashed line)

4.4 Marginal distributions

According to Figure 5, the models with non-homogeneous transition probabilities provide a better description of the marginal distribution of the wind direction than homogeneous models which have difficulties in reproducing the second mode of the distribution (associated to easterlies). The $\mathbf{NHMS-AR}_{(U,\Phi)}$ model seems to perform slightly better than the $\mathbf{NHMS-AR}_{(u,v)}$ model.

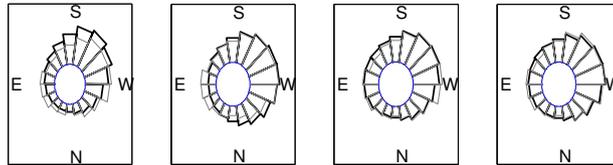


Figure 5: Rose plot of the marginal distribution of wind direction for the various models with $M = 4$ (resp. $M = 3$) regimes for Φ_t (resp. $\{u_t, v_t\}$) and order $s = 2$. From left to right: $\mathbf{HMS-AR}_{(u,v)}$, $\mathbf{NHMS-AR}_{(U,\Phi)}$, $\mathbf{NHMS-AR}_{(u,v)}$, $\mathbf{NHMS-AR}_{(U,\Phi)}$

The joint distribution of $\{u_t, v_t\}$ is globally well reproduced by both non-homogeneous models (see Figure 6). Simulated data exhibit two modes like in the original data. The modes seem to be slightly better located with the $\mathbf{NHMS-AR}_{(U,\Phi)}$ model. It may be due to the small differences in the non-homogeneous transition probabilities shown in Figure 4 with the $\mathbf{NHMS-AR}_{(U,\Phi)}$ model having a slightly higher probability of staying in regime 3 when the wind is blowing from the south-west. It may help to create two distinct modes at the correct locations.

Both models generate too much low wind and as a consequence fail to reproduce accurately the hole at origin. Similar lack of fit for the marginal distribution was observed on other datasets with MS-AR models. It seems to be especially sensible when the model is misspecified. This discrepancy may be reduced by developing alternative estimation methods which would give

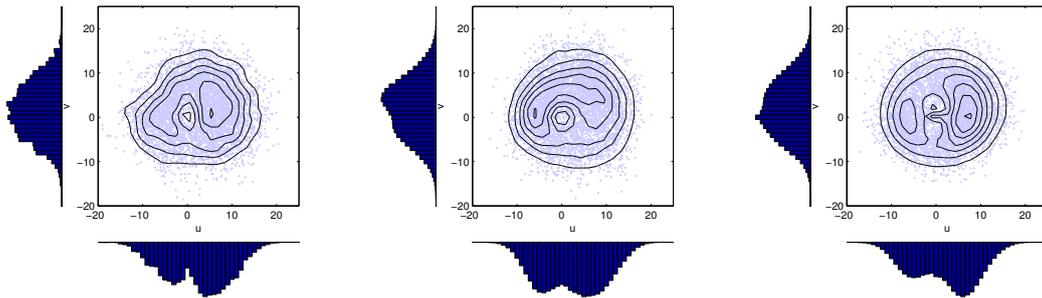


Figure 6: Joint distributions of $\{u_t, v_t\}$ for observed wind (left) and wind simulated with $\text{NHMS-AR}_{(u,v)}$ model (middle) and $\text{NHMS-AR}_{(U,\Phi)}$ model (right)

more importance to the stationary distribution of the process. This will be the topic of future research.

4.5 Dependence structure

All the models reproduce approximatively the first lags of autocorrelation function of $\{U_t\}$ (see Figure 7) and the circular autocorrelation of Φ (not shown) defined as (see [10])

$$\rho(h) = \frac{E[\cos(Y_0) \cos(Y_h)] + E[\sin(Y_0) \sin(Y_h)] - E[\sin(Y_0) \cos(Y_h)] - E[\cos(Y_0) \sin(Y_h)]}{E[\cos(Y_0)^2]E[\sin(Y_0)^2] - E[\sin(Y_0) \cos(Y_0)]^2} \quad (14)$$

for any positive integer h . To further validate the models, we have also plotted the various terms which appear in (14). According to Figure 8, the autocorrelation function of $\{\cos(\Phi_t)\}$ is generally better reproduced than the one of $\{\sin(\Phi_t)\}$. The empirical autocorrelation of $\{\sin(\Phi_t)\}$ has a more complex shape, with a quick decrease close to the origin and a bump around 4 days, than the one of $\{\cos(\Phi_t)\}$ which exhibits a more monotonic decrease. Figure 7 shows the cross-correlation function between the time series $\{\cos(Y_t)\}$ and $\{\sin(Y_t)\}$. The sample cross-correlation function is at its maximum value for a lag between 18 hours and 24 hours, with the time series $\{\sin(Y_t)\}$ being in advance of the time series $\{\cos(Y_t)\}$. This may be related to the fact that, for the location of interest, the wind direction tends to rotate more often clockwise than anti-clockwise between two successive time steps (see Figure 9). Note that the complex parametrization of the von Mises autoregressive models (see Section 2.4) permits to model rotation in a prevailing direction and significantly improves the boxplot shown on Figure 9 compared to models with real parametrization (not shown). One can also remark that the first order autoregressive matrices of the $\text{NHMS-AR}_{(u,v)}$ model have diagonal coefficients which are close to each other and out-diagonal coefficients which are almost opposed and thus may be interpreted as the product of rotation and dilatation matrices. Figure 9 shows however that they do not generate enough anticlockwise rotations.

The non-homogeneous models generally lead to a better description of the correlation functions compared to the homogeneous models. All the models lead to an underestimation of the empirical autocorrelations functions of the time series $\{\cos(Y_t)\}$ and $\{\sin(Y_t)\}$. Increasing the order s of the autoregressive models leads to a better description of the second order structure of the process but models of order $s \geq 3$ cannot reproduce the second mode of the marginal distribution and thus models of order $s = 2$ seem to provide a good compromise.

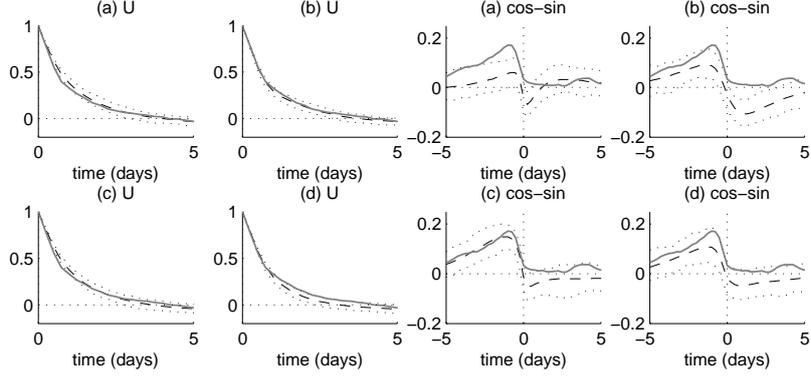


Figure 7: Correlation function of $\{U_t\}$ and cross-correlation functions between the time series $\{\cos(Y_t)\}$ and $\{\sin(Y_t)\}$ for the various models. The full grey line corresponds to the sample functions and the dashed line to the fitted model with a 95% prediction intervals (dotted line). The distributions for the fitted model was obtained by simulation. Time on the x-axis is expressed in days. (a): $\mathbf{HMS-AR}_{(u,v)}$, (b): $\mathbf{HMS-AR}_{(U,\Phi)}$, (c): $\mathbf{NHMS-AR}_{(u,v)}$, (d): $\mathbf{NHMS-AR}_{(U,\Phi)}$

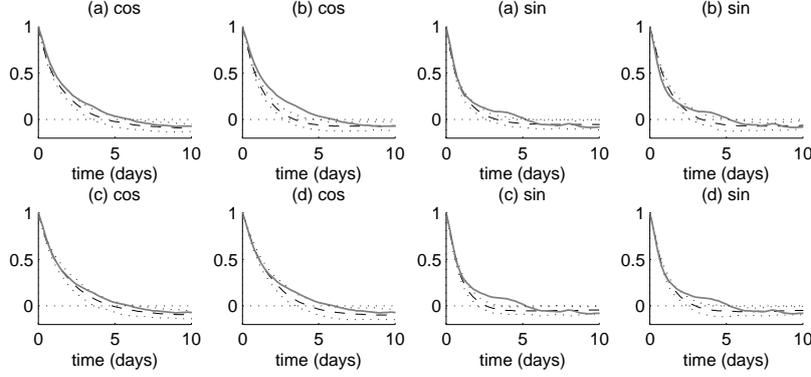


Figure 8: Autocorrelation functions of the time series $\{\cos(Y_t)\}$ and $\{\sin(Y_t)\}$ for the various models. The full grey line corresponds to the sample functions and the dashed line to the fitted model with a 95% prediction intervals (dotted line). The distributions for the fitted model was obtained by simulation. Time on the x-axis is expressed in days. (a): $\mathbf{HMS-AR}_{(u,v)}$, (b): $\mathbf{HMS-AR}_{(U,\Phi)}$, (c): $\mathbf{NHMS-AR}_{(u,v)}$, (d): $\mathbf{NHMS-AR}_{(U,\Phi)}$

5 Conclusion

In this paper we propose to model bivariate wind time series considering Cartesian coordinates on one hand and polar coordinates on the other hand. Both approaches have advantages. The $\{u_t, v_t\}$ model is easier to write and to fit since it is based on Gaussian distributions whereas a linear-circular model is required when considering polar coordinates. The $\{u_t, v_t\}$ model permits to globally well restore the second order structure observed on the data while the $\{U_t, \Phi_t\}$ model seems to give a better description of the marginal distributions. However, the differences between both models are slight.

Models with homogeneous and non-homogeneous latent Markov chains are compared. In non-homogeneous models, the transitions depend on the wind direction at the previous time. At the location of interest, wind is rotating more often clockwise but wind direction may also oscillate around two prevailing directions (northeast for anti-cyclonic conditions and southwest

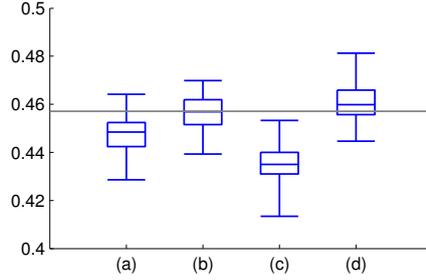


Figure 9: Frequency of anticlockwise rotations between successive observations for the various models with $M = 4$ regimes and autoregressive model of order $s = 2$. The grey line corresponds to the value obtained on the data (45.4 % of anticlockwise rotations against 54.6 % of clockwise rotations). The boxplots show the distributions for the fitted models. They were obtained by simulation (results based on 4400 time series of length 124). (a): **HMS-AR_(u,v)**, (b): **HMS-AR_(U,Φ)**, (c): **NHMS-AR_(u,v)**, (d): **NHMS-AR_(U,Φ)**

for cyclonic conditions). These features induce respectively some cycles which can be seen in the second order structure and modes in the marginal distribution. In broad outline, non-homogeneous transitions help the process to stay in the same weather regime when the wind direction is close to the prevailing directions and lead to sojourn duration in the regimes which are not geometric. In order to generate the prevailing rotation, it is necessary to command the wind direction to turn in the right direction inside the regimes. In $\{u_t, v_t\}$ models the rotations are reproduced by the autoregressive A matrices, but they are specified more naturally in **NHMS-EVM** model by using a complex parametrization of the von Mises autoregressive models.

The proposed models allow to generate wind time series with features very close to the main features of the observed time series. The introduction of the latent state allows to simulate the different time scale which are present in the data, with the autoregressive part describing the short-term fluctuations whereas the weather type, which lasts typically a few days, describes longer-term fluctuations and is related to large scale circulation. Another layer could be added to simulate shorter time scales for very local features. The model could also be extended to a space-time model in several ways. For this, it will probably be easier to work with the $\{u_t, v_t\}$ model based on Gaussian distributions which can naturally handle a space-time information. Then several strategies could be considered for the weather type process which could be local, with a different weather type at each site, or regional with a common weather type for the different locations. With the first strategy one has to deal with a space-time process of latent discrete variables and this is challenging from both a modeling and computational point of view. The second strategy is probably simpler to implement but requires some space-time homogeneity in the data. These and other related modeling issues are currently investigated.

Acknowledgments: We thank Julien Cattiaux (CNRS) for providing data and very helpful comments on this work.

References

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

- [2] P. Ailliot, V. Monbet, and M. Prevosto. An autoregressive model with time-varying coefficients for wind fields. *Environmetrics*, 17(2):107–117, 2006.
- [3] P. Ailliot and F. Pène. Consistency of the maximum likelihood estimate for non-homogeneous markov-switching models. *arXiv preprint arXiv:1306.2116*, 2014.
- [4] L.E. Baum, T. Petrie, G. Soules, and N. Weiss. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970.
- [5] J. Breckling. *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*, volume 61 of *Lecture Notes in Statistics*. Springer, Berlin, Germany, 1989.
- [6] 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.
- [7] O. Cappé, E. Moulines, and Rydén T. *Inference in hidden Markov models*. Springer-Verlag, New York, 2005.
- [8] J. Cattiaux, H. Douville, and Y. Peings. European temperatures in cmip5: origins of present-day biases and future uncertainties. *Climate Dynamics*, 41(11-12):2889–2907, 2013.
- [9] A.P. Dempster, Laird N.M., and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39(1):1–38, 1977.
- [10] N.I. Fisher and A. J. Lee. A correlation coefficient for circular data. *Biometrika*, 70:327–332, 1983.
- [11] N.I. Fisher and J. Lee. Time series analysis of circular data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 56:327–339, 1994.
- [12] M.D. Fraser, Y.S. Hsu, and Walker J.J. Identifiability of finite mixtures of von mises distributions. *The annals of statistics*, 9:1130–1131, 1981.
- [13] M. Fuentes, L. Chen, J.M. Davis, and G.M. Lackmann. Modeling and predicting complex space–time structures and patterns of coastal wind fields. *Environmetrics*, 16(5):449–464, 2005.
- [14] J.D. Hamilton. Analysis of time series subject to changes in regime. *Journal of Econometrics*, 45:39–70, 1990.
- [15] H. Holzmann, A. Munk, M. Suster, and W. Zucchini. Hidden markov models for circular and linear-circular time series. *Environmental and Ecological Statistics*, 13:325–347, 2006.
- [16] J.P. Hughes, Guttorp P., and S.P. Charles. A non-homogeneous hidden markov model for precipitation occurrence. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(1):15–30, 1999.
- [17] S. Kato. A markov process for circular data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72:655–672, 2010.

- [18] D. Maraun, F. Wetterhall, A.M. Ireson, R.E. Chandler, E.J. Kendon, M. Widmann, S. Brienen, H.W. Rust, T. Sauter, M. Themeßl, et al. Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, 48(3), 2010.
- [19] K.V. Mardia. *Statistics of directional data*. Academic press, New York, 1972.
- [20] 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.
- [21] P. Pinson, L.E.A. Christensen, H. Madsen, P.E. Sørensen, M.H. Donovan, and L.E Jensen. Regime-switching modelling of the fluctuations of offshore wind generation. *Journal of Wind Engineering and Industrial Aerodynamics*, 96(12):2327–2347, 2008.
- [22] R. Srikanthan and T.A. McMahon. Stochastic generation of annual, monthly and daily climate data: A review. *Hydrology and Earth System Sciences*, 5(4):653–670, 1999.
- [23] H. Tong. *Non-Linear Time Series: A Dynamical System Approach*. Oxford University Press, Oxford, U.K., 1990.
- [24] R. Vautard. Multiple weather regimes over the north atlantic: Analysis of precursors and successors. *Monthly Weather Review*, 118(10):2056–2081, 1990.
- [25] C.K. Wikle, R.F. Milliff, D. Nychka, and L.M. Berliner. Spatiotemporal hierarchical bayesian modeling tropical ocean surface winds. *Journal of the American Statistical Association*, 96(454):382–397, 2001.
- [26] W. Zucchini and P. Guttorp. A hidden Markov model for space-time precipitation. *Water Resources Research*, 27:1917–1923, 1991.
- [27] W. Zucchini and I. MacDonald. *Hidden Markov Models for Time Series: An Introduction Using R*. Number 110 in Monographs on statistics and applied probability. CRC Press, 2009.