Comparison of hidden and observed regime-switching autoregressive models for (u,v)-components of wind fields in the Northeast Atlantic

Julie Bessac, Pierre Ailliot, Julien Cattiaux, Valérie Monbet

Manuscript prepared for J. Name
with version 2015/04/24 7.83 Copernicus papers of the LaTeX class copernicus.cls.
Date: 4 January 2016

Comparison of hidden and observed regime-switching autoregressive models for (u,v)-components of wind fields in the Northeast Atlantic

Julie Bessac [1,2], Pierre Ailliot [3], Julien Cattiaux [4], and Valerie Monbet [1,5]

[1]Institut de Recherche Mathématiques de Rennes, UMR 6625, Université de Rennes 1, Rennes, France
[2]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, USA
[3]Laboratoire de Mathématiques de Bretagne Atlantique, UMR 6205, Université de Brest, Brest, France
[4]CNRM-GAME, UMR 3589, CNRS/Météo France, Toulouse, France
[5]INRIA Rennes, ASPI, Rennes, France

Correspondence to: Julie BESSAC (jbessac@anl.gov)

Abstract. Several multisite stochastic generators of zonal and meridional components of wind are proposed in this paper. A regime-switching framework is introduced to account for the alternation of intensity and variability that is observed on wind conditions due to the existence of different weather types. This modeling blocks time series into periods in which the series is described by a single model. The regime-switching is modeled by a discrete variable that can be introduced as a latent (or hidden) variable or as an observed variable. In the latter case a clustering algorithm is used before fitting the model to extract the regime. Conditionally to the regimes, the observed wind conditions are assumed to evolve as a linear Gaussian vector autoregressive (VAR) model. Various questions are explored, such as the modeling of the regime in a multisite context, the extraction of relevant clusterings from extra-variables or from the local wind data, and the link between weather types extracted from wind data and large-scale weather regimes derived from a descriptor of the atmospheric circulation. We also discuss relative advantages of hidden and observed regime-switching models. For artificial stochastic generation of wind sequences, we show that the proposed models reproduce the average space-time motions of wind conditions; and we highlight the advantage of regime-switching models in reproducing the alternation of intensity and variability in wind conditions.

1 Introduction

In this section, we present the context of our work and then the data used to compare the proposed Markov-switching autoregressive models.

1.1 Introduction

Stochastic weather generators have been used to generate artificial sequences of small-scale meteorological data with statistical properties similar to the dataset used for calibration. Various wind condition generators at a single site have been proposed in the literature; see (Brown, Katz, and Murphy, 1984; Flecher, Naveau, Allard, and Brisson, 2010; Ailliot and Monbet, 2012). However, few models have been introduced in a multisite context (Haslett and Raftery, 1989; Bessac, Ailliot, and Monbet, 2015). Artificial sequences of wind conditions provided by stochastic weather generators enable assessment risks in impact studies; see, for instance, (Hofmann and Sperstad, 2013). Here we propose a multisite generator for Cartesian components of surface wind. As far as we know, only a few models have been proposed to simulate time series of Cartesian coordinates of wind $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$ (Hering, Kazor, and Kleiber, 2015; Hering and Genton, 2010; Ailliot, Monbet, and Prevosto, 2006; Wikle, Milliff, Nychka, and Berliner, 2001; Fuentes, Chen, Davis, and Lackmann, 2005). Except in (Hering, Kazor, and Kleiber, 2015), these models are designed for short-term wind prediction and not for the generation of artificial conditions of $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$. Consequently they are not focused on reproducing the same statistics we are interested in, namely, the marginal distribution of $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$ and its spatiotemporal dynamics. In (Hering, Kazor, and Kleiber, 2015) a stochastic generator for multiple temporal and spatial scales is proposed. The proposed Markov-switching vector autoregressive model enables reproduction of many spatial and temporal features; however complex dependencies between intensity and direction remain hard to model.

In the Northeast Atlantic, the spatiotemporal dynamics of the wind field is complex. This area is under the influence of an unstable atmospheric jet stream, whose large-scale fluctuations induce local alternations between periods with high wind intensity and strong temporal variability, and less intense and variable periods. Scientists have proposed describing the North-Atlantic atmospheric dynamics through a finite number of preferred states, namely, weather regimes or weather types (Vautard, 1990). However, introducing regime-switching in the modeling of local wind, as we propose in this paper, enables us to better reproduce the spatiotemporal characteristics observed in the wind data. In practice, describing a time series by regimes involves a partitioning into time periods in which the series is homogeneous and can be described by a single model. In this paper, we propose various vector autoregressive (VAR) models with regime-switching. One of the challenges is to achieve a regime-switching that is physically consistent and that enables appropriately describing the local observation by a VAR model. To this end, we introduce several frameworks of regime-switching and compare them in terms of simulation of wind data.

Depending on the availability of good descriptors of the current weather state, regime-switching can be introduced with either observed or latent regimes. Regimes are said to be observed when they are identified a priori, before the modeling of the local dynamics. In this case, clustering methods are run on adequate variables to obtain relevant regimes: either the local variables or extra-variables characterizing the large-scale weather situation, such as descriptors of the large-scale atmospheric

circulation (Bardossy and Plate, 1992; Wilson, Lettenmaier, and Skyllingstad, 1992) or variables enabling the separation into dry and wet states (Richardson, 1981; Flecher, Naveau, Allard, and Brisson, 2010). For wind models, the wind direction can be considered since it is a good descriptor of synoptic conditions. In (Gneiting, Larson, Westrick, Genton, and Aldrich, 2006), the wind direction is used both to extract regimes and to parameterize of the predictive distribution. In this paper, we propose a priori clusterings based on both large-scale and local variables.

When the regimes are said to be latent, they are introduced as a hidden variable in the model. This framework is more complex from a statistical point of view and the conditional distribution of wind given the regime has to be simple and tractable. Hidden Markov models (HMMs) have been widely used for meteorological data (Zucchini and Guttorp, 1991; Hughes, Guttorp, and Charles, 1999; Thompson, Thomson, and Zheng, 2007). Hidden Markov-switching autoregressive (MS-AR) models are a generalization of HMMs allowing temporal dynamics within the regimes (Hamilton, 1989). Models with regime-switching improve the modeling of wind intensity time series with classical autoregressive–moving-average (ARMA) models; see (Ailliot and Monbet, 2012), where the wind speed is modeled at one site. Here we propose a hidden MS-AR model and compare it with several models with observed regime-switching.

To the best of our knowledge, no comparison between observed and latent regime-switching has been proposed in the field of stochastic generators of wind conditions. In (Pinson, Christensen, Madsen, Sorensen, Donovan, and Jensen, 2008), a comparison is presented in terms of wind prediction between models with hidden regimes and models driven by observed regimes. In this work, we compare both kinds of models in a simulation framework.

In the multisite context, the regime can be either common to all sites (i.e., scalar; see (Ailliot, Thompson, and Thomson, 2009)) or introduced as a site-specific regime (Wilks, 1998; Kleiber, Katz, and Rajagopalan, 2012; Khalili, Leconte, and Brissette, 2007; Thompson, Thomson, and Zheng, 2007), which enables one to account for a wide range of space-time dependencies. However, a site-specific regime appears to be computationally challenging (Wilks, 1998). We will show that the choice of a regional regime is reasonable when a homogeneous area is selected.

The paper is organized as follows. MS-AR models are introduced in Section 2, and their inference is described in cases of both observed and latent regime-switching. The question of a regional regime is addressed in Section 3. In Section 4, we introduce and discuss different sets of a priori regimes obtained by clustering. In Sections 6 and 7, respectively we discuss the advantages of the proposed models and highlight the differences between observed and latent regime-switching models.

1.2 Wind data

The data under study are zonal (west-east) and meridional (north-south) surface wind components $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$ at 10 meters above sea level extracted from the ERA-Interim dataset produced by the Eu-

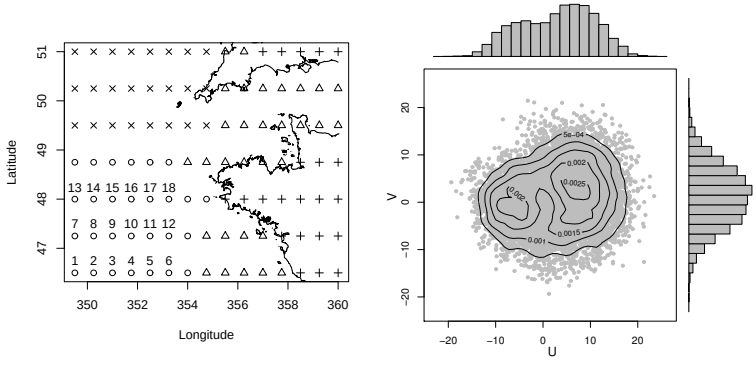


Figure 1. Left: Spatial hierarchical clustering of the moving variance associated with wind speed with four clusters (symbols). Right: Joint and marginal distribution of $\{u_t, v_t\}$ at the central location 10; contour lines of the estimated joint density.

ropean Center of Medium-range Weather Forecast (ECMWF). It can be freely downloaded from the URL `http://data.ecmwf.int/data/` and used for scientific purposes.

We focus on gridded locations between latitudes $46.5°$N and $48°$N and longitudes $6.75°$W and $10.5°$W (15×7 grid points; see Figure 1). The dataset we have extracted consists of 32 December-January blocks of wind data from December 1979 to January 2011 picked every 6 hours. Further, the statistical inference is based on the assumption that the 32 December-January blocks of wind components are 32 independent realizations of the same stationary process, a reasonable assumption given the strong interannual variability of the wintertime atmospheric dynamics at such a local scale. The training dataset is then composed of 32 independent blocks and each block has 4×62 observations. In order to study the relevance of using common regimes for all the locations, a spatial hierarchical clustering has been used to choose a homogeneous area (see Figure 1). The clustering is run on the process of moving standard deviation of wind speed, which is described more precisely in Section 6. This process is a good descriptor of the temporal characteristics of wind time series (see Figure 4), and it is computed as the standard deviation of wind speed over nine consecutive time steps (i.e., two days). The dendogram associated with the clustering suggests the use of four clusters that are depicted on Figure 1. These four clusters are likely to be divided into an inland cluster ($+$), an intermediate cluster between ocean and land ($\triangle$), a cluster corresponding to flows that propagate into the Bay of Biscay ($\circ$), and a cluster for flows that propagate toward northern Europe ($\times$).

Components $\{u_t\}$ and $\{v_t\}$ admit a complex relationship, as partially reflected by the joint distribution of $\{u_t, v_t\}$ (Figure 1). The margin of $\{u_t\}$ reveals two separate modes, whereas that of $\{v_t\}$ does not exhibit a clear bimodality. The contour lines show that the density is low around the point (0,0). It indicates that the transitions between the two modes of each component are not realized through a vanishing of the field but rather through a rotation of the field. The following transfor-

mation is used on both components $\{u_t\}$ and $\{v_t\}$. This transformation with $\alpha > 1$ facilitates the modeling of the bimodality:

$$\begin{cases} \tilde{u}_t = & U_t^{\alpha} \cos(\Phi_t) \\ \tilde{v}_t = & U_t^{\alpha} \sin(\Phi_t), \end{cases} \tag{1}$$

where $\{U_t\}$ and $\{\Phi_t\}$ respectively denote wind speed and wind direction. In practice, α is chosen empirically equal to 1.5. This transformation has proven helpful in modeling the distribution of $\{u_t, v_t\}$ in (Ailliot, Bessac, Monbet, and Pene, 2015).

2 Markov-switching vector autoregressive models

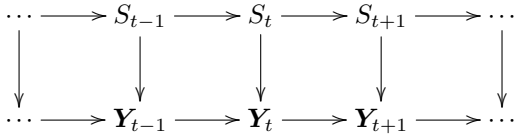
In this section, we introduce the proposed models and discuss their parameter estimation in cases of both observed and latent regimes.

2.1 The models

In this paper, we consider the following class of models. Let S_t be a discrete Markov chain with values in $\{1, ..., M\}$ describing the current weather type as a function of time t. Conditionally to the weather type, the observed wind conditions are modeled as a vector autoregressive model. Given the current value of S_t, the observation $\boldsymbol{Y}_t$ is written as

$$\boldsymbol{Y}_t = \boldsymbol{A}_0^{(S_t)} + \boldsymbol{A}_1^{(S_t)} \boldsymbol{Y}_{t-1} + \boldsymbol{A}_2^{(S_t)} \boldsymbol{Y}_{t-2} + ... + \boldsymbol{A}_p^{(S_t)} \boldsymbol{Y}_{t-p} + (\boldsymbol{\Sigma}^{(S_t)})^{-1/2} \boldsymbol{\epsilon}_t. \tag{2}$$

$\boldsymbol{Y} \in \mathrm{R}^{2K}$ represents the observed power-transformed wind components $\{u_t, v_t\}$ at the K locations, given by the system (1). For $i \in \{1, ..., M\}$, $\boldsymbol{A}_0^{(i)}$ is a $2K$-dimensional vector, $\boldsymbol{A}_1^{(i)}, ..., \boldsymbol{A}_p^{(i)}, \boldsymbol{\Sigma}^{(i)}$ are $2K \times 2K$-matrices, and $\boldsymbol{\epsilon}$ is a Gaussian white noise of dimension $2K$. Conditional independencies between S and $\boldsymbol{Y}$ are displayed on the following directed acyclic graph (DAG) for $p = 1$ (see (Durand, 2003) for additional information about DAGs):



In this model, the regime S can be latent or observed; both cases are discussed, respectively, in Sections 3 and 4. The parameter estimation of the model can be performed by maximum likelihood but in a different way in each framework.

For both kind of models, covariates can be included. The easiest way is to include them in the intercept parameter $\boldsymbol{A}_0$ or in transitions between regimes. Transitions between regimes can be parametrized with a covariate (when regimes are latent, a parameterization with an extra covariate is given in (Hughes and Guttorp, 1994) and with the studied variable in (Ailliot, Bessac, Monbet,

and Pene, 2015) and in (Vrac, Stein, and Hayhoe, 2007) when regimes are defined a priori). In the context of multisite models, the choice of the covariate of non-homogeneous transitions is delicate. We do not discuss this topic here and consider only homogeneous transition models.

To avoid overparameterization of the conditional models, we first work with a reduced dataset. In the following all the proposed models will be fitted on the subset of sites (1,6,10,13,18), the extension to a wider region being left for future studies.

2.2 Estimation by maximum likelihood

First, let us suppose that the complete set of observations $(\boldsymbol{y}_1,...\boldsymbol{y}_T,s_1,...s_T)$ is available, which is the case in Section 4. Assume that s_0, $\boldsymbol{y}_{-1}$ and $\boldsymbol{y}_0$ are observed. Then the complete log-likelihood, associated with an autoregressive order $p=2$ (we choose $p=2$ according to a previous work (Ailliot, Bessac, Monbet, and Pene, 2015)), is written as

$$\begin{aligned}\log(\mathcal{L}(\theta;\boldsymbol{y}_1,...\boldsymbol{y}_T,s_1,...s_T|\boldsymbol{y}_{-1},\boldsymbol{y}_0,s_0)) &= \log(\mathcal{L}(\theta^{(\boldsymbol{Y})};\boldsymbol{y}_1^T|\boldsymbol{y}_{-1},\boldsymbol{y}_0,s_0^T)) \\ &\quad +\log(\mathcal{L}(\theta^{(S)};s_1^T|\boldsymbol{y}_{-1},\boldsymbol{y}_0,s_0)), \qquad (3)\end{aligned}$$

where $\theta=(\theta^{(S)},\theta^{(\boldsymbol{Y})})$. $\theta^{(\boldsymbol{Y})}$ corresponds to the parameters of the VAR models, $\theta^{(S)}=\boldsymbol{\Pi}=(\pi_{i,j})_{i,j=1,\cdots,M}$ the transition matrix $\boldsymbol{\Pi}$ of the Markov chain S, and $\boldsymbol{y}_1^T=(\boldsymbol{y}_1,...,\boldsymbol{y}_T)$. Let us denote $n_{i,j}$ the number of occurrences of the event $\{(S_t,S_{t+1})=(i,j)\}$ for $t\in\{1,...,T-1\}$, $n_{i,.}=\sum_{j=1}^M n_{i,j}$ and $n_i=n_{i,.}+\delta_{\{s_T=i\}}$, where δ is the Kronecker symbol, the total number of occurrences of the regime i:

$$\begin{aligned}&\log(\mathcal{L}(\theta^{(\boldsymbol{Y})};\boldsymbol{y}_1,...,\boldsymbol{y}_T|\boldsymbol{y}_{-1},\boldsymbol{y}_0,s_0^T)) \\ =&\sum_{t=1}^{T}\log(p(\boldsymbol{y}_t|\boldsymbol{y}_{t-1},\boldsymbol{y}_{t-2},s_t)) \\ =&\sum_{i=1}^{M}\sum_{t\in\{t|s_t=i\}}\log(p(\boldsymbol{y}_t|\boldsymbol{y}_{t-1},\boldsymbol{y}_{t-2},s_t)) \\ =&\sum_{i=1}^{M}n_i(-\tfrac{d}{2}\log(2\pi)-\tfrac{1}{2}\log(\det(\boldsymbol{\Sigma}^{(i)}))-\sum_{t\in\{t|s_t=i\}}\tfrac{1}{2}\boldsymbol{e}_t'(\boldsymbol{\Sigma}^{(i)})^{-1}\boldsymbol{e}_t,\end{aligned}$$

where $\boldsymbol{e}_t=(\boldsymbol{y}_t-\boldsymbol{A}_0^{(i)}-\boldsymbol{A}_1^{(i)}\boldsymbol{y}_{t-1}-\boldsymbol{A}_2^{(i)}\boldsymbol{y}_{t-2})$.

For each $i\in\{1,...,M\}$, each function

$$\theta^{(\boldsymbol{Y},i)} \qquad \to n_i(-\tfrac{d}{2}\log(2\pi)-\tfrac{1}{2}\log(\det(\boldsymbol{\Sigma}^{(i)}))-\sum_{t\in\{t|s_t=i\}}\tfrac{1}{2}\boldsymbol{e}_t'(\boldsymbol{\Sigma}^{(i)})^{-1}\boldsymbol{e}_t$$

can be maximized separately, where $\theta^{(\boldsymbol{Y},i)}=(\boldsymbol{A}_0^{(i)},\boldsymbol{A}_1^{(i)},\boldsymbol{A}_2^{(i)},\boldsymbol{\Sigma}^{(i)})$. The optimal estimates of $\boldsymbol{A}_1^{(i)}$ and $\boldsymbol{A}_2^{(i)}$ are computed by writing the VAR(2) model as a VAR(1): for all $t\in\{t|s_t=i\}$,

$$\begin{pmatrix}\boldsymbol{Y}_t\\ \boldsymbol{Y}_{t-1}\end{pmatrix}=\begin{pmatrix}\boldsymbol{A}_1^{(i)} & \boldsymbol{A}_2^{(i)}\\ \mathbf{Id}_{\boldsymbol{K}} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\boldsymbol{Y}_{t-1}\\ \boldsymbol{Y}_{t-2}\end{pmatrix}+\begin{pmatrix}\boldsymbol{\epsilon}_t\\ \mathbf{0}\end{pmatrix},$$

where Id_K is the $K \times K$-identity matrix. Let us write $\boldsymbol{A}^{(i)} = \begin{pmatrix} \boldsymbol{A}_1^{(i)} & \boldsymbol{A}_2^{(i)} \\ \mathbf{Id_K} & \mathbf{0} \end{pmatrix}$ and $\boldsymbol{Z}_t = \begin{pmatrix} \boldsymbol{Y}_t \\ \boldsymbol{Y}_{t-1} \end{pmatrix}$;

expressions of $\hat{\boldsymbol{A}}_1^{(i)}$ and $\hat{\boldsymbol{A}}_2^{(i)}$ are extracted from the estimate

$$\hat{\boldsymbol{A}^{(i)}} = \Big(\sum_{t \in \{t | s_t = i\}} \boldsymbol{Z}_t \boldsymbol{Z}_{t-1}^{'} \Big) \Big(\sum_{t \in \{t | s_t = i\}} \boldsymbol{Z}_{t-1} \boldsymbol{Z}_{t-1}^{'} \Big)^{-1}. \tag{4}$$

The other optimal estimates are

$$\hat{\boldsymbol{A}}_0^{(i)} = (\mathbf{Id_K} - \hat{\boldsymbol{A}}_1^{(i)} - \hat{\boldsymbol{A}}_2^{(i)}) \hat{\boldsymbol{\mu}}^{(i)}, \tag{5}$$

where $\hat{\boldsymbol{\mu}}^{(i)} = \dfrac{1}{n_i} \displaystyle\sum_{t \in \{t | s_t = i\}} \boldsymbol{y}_t$ is the empirical mean of $\boldsymbol{Y}$ in regime i and

$$\hat{\boldsymbol{\Sigma}}^{(i)} = \frac{1}{n_i} \sum_{t \in \{t | s_t = i\}} \hat{\boldsymbol{e}}_t \hat{\boldsymbol{e}}_t^{'}, \tag{6}$$

$\hat{\boldsymbol{\Sigma}}^{(i)}$ is the empirical variance of the empirical residuals defined as $\hat{\boldsymbol{e}}_t = (\boldsymbol{y}_t - \hat{\boldsymbol{A}}_0^{(i)} - \hat{\boldsymbol{A}}_1^{(i)} \boldsymbol{y}_{t-1} - \hat{\boldsymbol{A}}_2^{(i)} \boldsymbol{y}_{t-2})$.

Concerning the Markov chain S,

$$\log(\mathcal{L}(\theta^{(S)}; s_1, ..., s_T | \boldsymbol{y}_{-1}, \boldsymbol{y}_0, s_0)) \quad = \quad \sum_{i,j=1}^{M} n_{i,j} \log(\pi_{i,j}),$$

the associated maximum likelihood estimator is

$$\hat{\pi}_{i,j} = \frac{n_{i,j}}{n_{i,.}}.$$

When observations only of the process $\boldsymbol{Y}$ are available and the realizations of S are not given a priori, as in Section 3, one inference method is to use the expectation-maximization (EM) algorithm, which is commonly run to estimate the parameters of models with latent variables by maximum likelihood. Since S is not observed, the EM algorithm aims at maximizing the incomplete log-likelihood function based on the observations $\boldsymbol{Y}$:

$$\theta \to \mathrm{E}_\theta(\log(\mathcal{L}(\theta; \boldsymbol{Y}_1, ..., \boldsymbol{Y}_T, S_1, ..., S_T)) | \boldsymbol{Y}_{-1}^T = \boldsymbol{y}_{-1}^T, S_0 = s_0).$$

It is proven that through the iterations of the algorithm, a convergent sequence of approximation of the maximum likelihood estimator of θ is computed.

EM algorithm cycles through two steps: the expectation step and the maximization step (Wu, 1983; Dempster, M., and Rubin, 1977). The E-step is performed through forward-backward recursions (see (Hamilton, 1990) for hidden MS-AR models) that enable one to compute the smoothing probabilities $P(S_t | \boldsymbol{Y}_{-1}^T = \boldsymbol{y}_{-1}^T, S_0 = s_0)$. At the M-step, optimal expressions of parameters of $\theta^{(\boldsymbol{Y})}$, given in (4), (5), and (6), are used. In each regime i, however, each observation $\boldsymbol{y}_t$ is weighted by the probability $P(S_t = i | \boldsymbol{Y}_{-1}^T = \boldsymbol{y}_{-1}^T, S_0 = s_0)$, for instance,

$$\hat{\boldsymbol{\mu}}^{(i)} = \frac{1}{\sum_{t=1}^T P(S_t = i | \boldsymbol{Y}_{-1}^T = \boldsymbol{y}_{-1}^T, S_0 = s_0)} \sum_{t=1}^{T} P(S_t = i | \boldsymbol{Y}_{-1}^T = \boldsymbol{y}_{-1}^T, S_0 = s_0) \boldsymbol{y}_t.$$

The transition matrix is estimated from quantities $P(\{S_t = i, S_{t+1} = j\}|\boldsymbol{Y}_{-1}^T = \boldsymbol{y}_{-1}^T, S_0 = s_0)$ that are derived at the E-step.

In this paper, we use AP-MS-VAR$_{\mathrm{C}}$ to denote the a priori regime-switching model associated with the clustering C, and we use H-MS-VAR to denote the hidden regime-switching model.

3 Regime definition in a multisite context

When the current weather state is not estimated a priori, it is introduced as a latent variable. Hidden regime-switching models have been used in various fields; see (Zucchini and MacDonald, 2009) for a wide range of applications of hidden Markov models. In a previous work (Ailliot, Bessac, Monbet, and Pene, 2015) a single-site model for $\{u_t, v_t\}$ was proposed, the proposed hidden Markov-switching autoregressive model reveals good qualities to describe both marginal and joint distributions of $\{u_t, v_t\}$ as well as the temporal dynamics of the wind at one location. In this paper we propose an extension of this model, when the process $\{u_t, v_t\}$ is multi-site. In a multi-site context, the regime can be site-specific or common to all stations.

Here, the assumption of a common regional regime is investigated, and we show that this assumption is acceptable when the considered area is homogeneous. The homogeneous single-site MS-AR model introduced in (Ailliot, Bessac, Monbet, and Pene, 2015) for $\{u_t, v_t\}$ with $M = 3$ regimes and an autoregressive order $p = 2$ has been fitted at each site. The most likely regimes associated with the data are extracted from the estimation procedure of H-MS-VAR models described in the previous section. At each time, the regime corresponds to $\underset{j\in\{1,\cdots,M\}}{\arg\max}\ P(S_t = j|\boldsymbol{Y}_{-1}^T = \boldsymbol{y}_{-1}^T, S_0 = s_0)$, see (Zucchini and MacDonald, 2009). In order to properly compare the regimes, they are ordered according to the increasing value of the determinant of the matrix $\boldsymbol{\Sigma}^{(i)}$. The intuition for sorting regimes according the determinant of $\boldsymbol{\Sigma}^{(i)}$ is that we expect the innovations to be more volatile, and consequently $\boldsymbol{\Sigma}^{(i)}$ having greater eigenvalues, in cyclonic weather regimes. Conversely, we expect to observe innovations more persistent in time in calm weather regimes, this is associated with smaller eigenvalues of $\boldsymbol{\Sigma}^{(i)}$. The spatiotemporal coherence of the regimes of each of the 18 sites is checked and reveals a strong homogeneity that motivates using a regional regime in this area.

The sequences of regimes are compared in Figure 2, time series of a posteriori regimes and wind speed are depicted. The last two regimes are less coherent from one site to another. This effect is partly explained by the fact that these regimes are less persistent in time, especially the third one (see Table 1). Moreover, we can notice an eastward propagation in wind events, the darkest regimes being often observed at western stations (station 1) prior to eastern sites (10 and 18). The bottom panel of the Figure 2, which depicts the sequences of regimes associated with the model fitted on the set of all locations with a common regime to all locations, reveals that this regional regime is coherent with the local ones, although it is less persistent. Indeed, when fitting the model to several

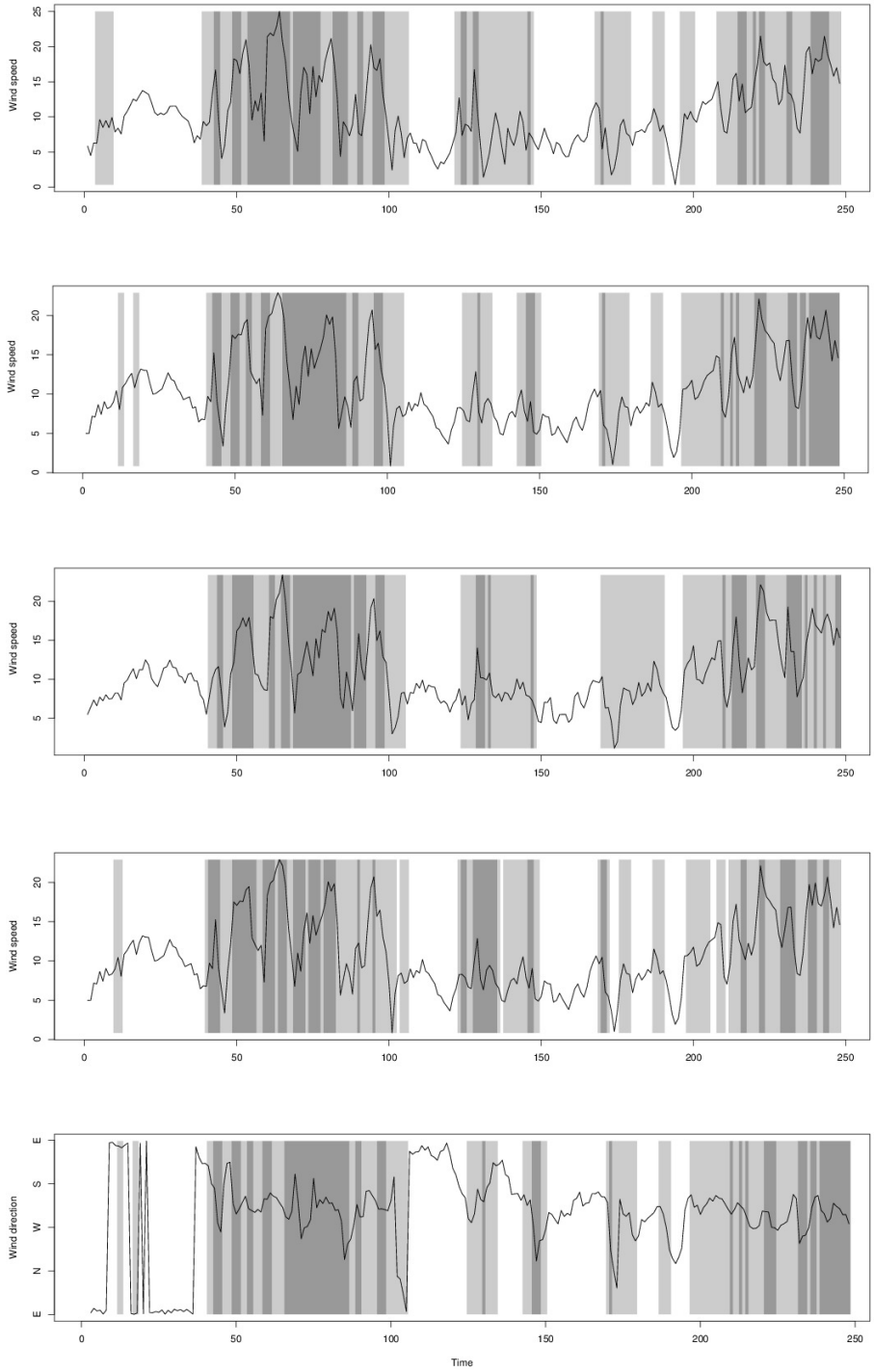


Figure 2. Time series of wind speed in January 2012 and a posteriori regimes from the fitting of a H-MS-VAR. The lighter is the grey; the smaller is the determinant of $\mathbf{\Sigma}^{(i)}$. From top to bottom: sites 1, 10, and 18 when the model is fitted at a single location, fourth panel from the top: extracted regimes when the model is fitted at the 5 locations (1,6,10,13,18). Bottom panel: wind direction and regimes at site 10.

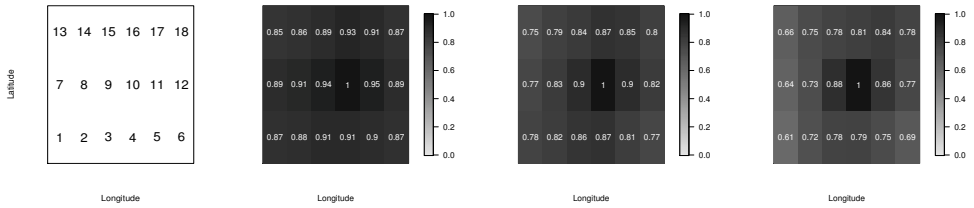


Figure 3. Most left panel: matrix with the number of the station is printed; then from left to right, conditional probabilities of occurrence of regime $i = 1, 2, 3$ at all sites conditional to the simultaneous occurrence of the same regime at site 10; on each pixel, the value of the conditional probability is plotted.

Table 1. Parameter values obtained when fitting a H-MS-VAR at the different sites: diagonal of the transition matrix $\mathbf{\Pi}$, coefficients of the autoregressive model in each regime, and logarithm of the determinant of $\mathbf{\Sigma}^{(i)}$.

	Diagonal of $\mathbf{\Pi}$			AR Coefficients ($\boldsymbol{A}_1^{(i)}(1,1), \boldsymbol{A}_1^{(i)}(2,2)$)			$\log(\det(\mathbf{\Sigma}^{(i)}))$		
Site \ Regime	R1	R2	R3	R1	R2	R3	R1	R2	R3
Site 1	0.93	0.83	0.64	(1.27,1.16)	(1.15,1.3)	(0.62,0.63)	5.62	8.87	11.96
Site 6	0.92	0.83	0.71	(1.27,1.02)	(1.2,1.28)	(0.61,0.72)	5.55	8.59	11.79
Site 10	0.93	0.84	0.74	(1.25,1.19)	(1.17,1.27)	(0.74,0.71)	5.55	8.67	11.79
Site 13	0.93	0.81	0.64	(1.22,1.24)	(1.17,1.25)	(0.65,0.65)	5.77	9	11.96
Site 18	0.93	0.83	0.73	(1.26,1.12)	(1.17,1.25)	(0.67,0.68)	5.72	8.73	11.83

stations, the regime has to embed some spatial heterogeneity that is likely to decrease the temporal persistence.

In Figure 3, probabilities of occurrence of a given regime conditional to the simultaneous occurrence of the same regime at site 10 are depicted for all sites. In each picture, conditional probabilities should be compared with the reference value given at location 10, which is 1 by construction. The first regime has the best spatial coherence; and the third regime, which is the least persistent regime, is less coherent spatially. The ranges of values of these probabilities indicate a satisfying consistency between the regimes across sites. At each site, the physical interpretation of each regime is similar. Indeed, the first regime corresponds mainly to anticyclonic conditions with easterly winds and a slowly varying intensity (the variance of the innovation of the AR model is lower than in the two other regimes, and the first AR coefficient is larger; see Table 1). The two other regimes correspond to cyclonic conditions with westerly winds and a higher temporal variability in the intensity (see Figure 4). These two regimes are discriminated mainly by the temporal variability, which is higher in the third regime. Moreover the wind direction, not depicted here, slightly differs: from south-westerlies in the second regime to northwesterlies in the third regime. In Figure 4, we can notice that wind conditions with weak temporal variability observed in the first regime are associated with weak values of the moving mean and variance processes, whereas more volatile periods in the second and third regimes are characterized by higher values of moving mean and variance. To the best of our

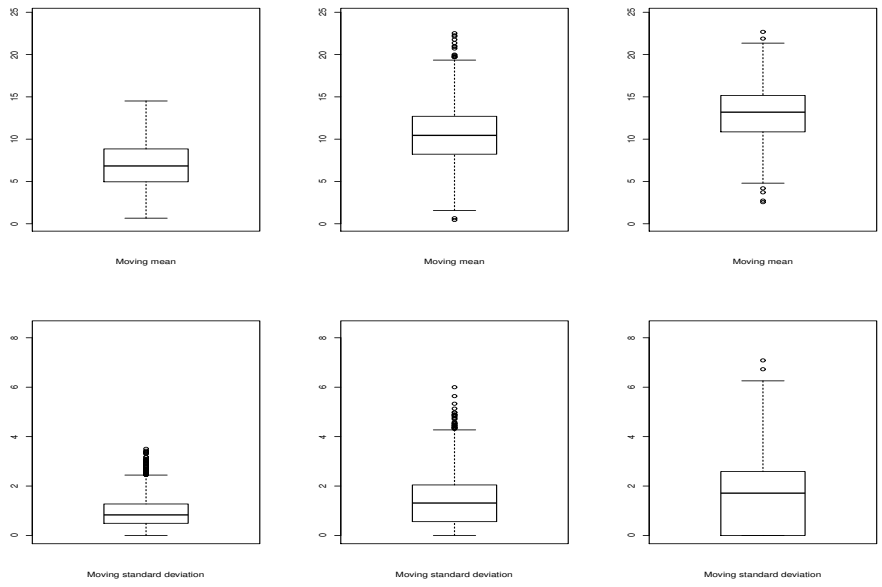


Figure 4. Top panel: moving mean of wind speed computed on two days intervals (nine time steps) for each regime of the H-MS-VAR model fitted at site 10. Bottom panel: same for moving standard deviation.

knowledge, few statistics enable us to characterize the alternation associated with regime-switching. These two processes of moving mean and standard deviation enable to characterize the alternation of variability associated with the observed regime-switching and will be used in the following sections.

Coefficients of the autoregressive process $\boldsymbol{Y}$ in each regime and the transition matrix at each site are comparable and spatially coherent (see Table 1). Other criteria such as the average field of $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$ in each regime and distribution of $\{\Phi_t\}$ in each regime were also explored and suggest similarities between regimes at all locations.

The assumption of a regional regime seems appropriate in the considered area and is thus kept for the modeling of the multisite wind in the following.

4 Observed regime-switching autoregressive models

Conversely to the previous section, one may derive the regimes separately from the fitting of the conditional model. For such a priori regime-switching models, the derivation of observed regimes can be done with appropriate clustering methods. We seek weather states that are distinct one from the other and in which the data are homogeneous. Clustering can be run either on the local variables under study or on extra-variables: the former leads to weather states that are more appropriate to the local data, while the latter can provide more meteorologically consistent regimes for example with

more information about the large-scale situation. In this subsection, we propose three clusterings, which differ by the clustering method and/or by the variables used to derive the a priori regimes.

4.1 Derivation of observed regimes from extra-variables: $\mathbf{C}_{Z500}$

As a first clustering, we use a classification into four large-scale weather regimes that is commonly used in climate studies to characterize the wintertime atmospheric dynamics over the North Atlantic / European sector ((Michelangeli, Vautard, and Legras, 1995; Cassou, 2008; Najac, 2008)). These regimes can be described as follows:

- The positive phase of the North-Atlantic Oscillation (hereafter NAO+), characterized by a strengthening of both the Azores High and the Islandic Low, which reinforces the westerlies
- The negative phase of the NAO (NAO−), its symmetrical counterpart
- The Scandinavian blocking (BL), characterized by a strong anticyclone over northern Europe able to totally block the westerly flow over western Europe,
- The Atlantic Ridge (AR), characterized by a strong west-east pressure dipole bringing polar air masses over western Europe

At the local scale of our area of study, these regimes are respectively associated with strong south-westerly flows (NAO+), weak westerly flows (NAO-), stable southerly or easterly flows (BL) and northerly flows (AR).

To derive these regimes, we use the same methodology as in (Cattiaux, Douville, and Peings, 2013). We perform a k-means clustering on the 3,607 daily-mean maps of 500 mb geopotential height (Z500) anomalies (i.e., mean-corrected fields) over the North Atlantic / European sector (90°W-30°E / 20-80°N) corresponding to days of December, January, and February 1981–2010. Daily Z500 data are downloaded from the ERA-Interim archive. In order to reduce the computational time, the k-means algorithm is performed on the first ten principal components (PCs) of the Z500 anomalies time series. These PCs are time series corresponding to the projections of the Z500 anomalies onto the empirical orthogonal functions (EOFs), which are eigenvectors of the spatial covariance matrix of the Z500 field. Such a decomposition enables extraction of the main modes of variability of the spatiotemporal process; here, the first ten EOFs explain 90% of the total variance. Eventually, the obtained daily classification is converted to a $4\times$daily classification by repeating the same regime for the four time steps of each day, a reasonable approach given the smoothness of the Z500 both in time and space. In the following, we denote this clustering C_{Z500}.

4.2 Derivation of observed regimes from the local variables: $\mathbf{C}_{EOF(u,v)}$ and $\mathbf{C}_{Diff(u,v)}$

To derive observed regimes from local wind variables, one can first use a k-means clustering procedure similar to the one used for C_{Z500}. However, while C_{Z500} provides persistent regimes in which

the conditional model satisfyingly describes $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$, local regimes resulting from such a k-means clustering are not persistent enough to reliably estimate the conditional VAR model. Consequently, in this subsection, we perform the local clustering via a hidden Markov model with Gaussian probability of emission.

The hidden structure of the Markov chain provides more stable regimes than with a k-means clustering. It corresponds to an H-MS-VAR model with VAR models of order $p = 0$. The EM algorithm is used to process the clustering, and the number of regimes is chosen at three. This number provides the most physically relevant local regimes; a greater number of regimes indeed leads to less discriminative regimes in terms of local wind conditions (not shown).

Then two sets of descriptors of the data (i.e. local variables) are proposed. The first partition, denoted $\mathrm{C}_{EOF(u,v)}$, is obtained by clustering the time series associated with the first two EOFs of the anomalies of $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$, which explain 94% of the total variance. The second partition involves descriptors of the conditional distribution of $p(\boldsymbol{Y}_t|\boldsymbol{Y}_{t-1})$, in order to find a clustering that may be better adapted to the description of the conditional distribution by an autoregressive model. A simplified way to describe the dynamics is to consider the bivariate process $\{\boldsymbol{u}_t - \boldsymbol{u}_{t-1}, \boldsymbol{v}_t - \boldsymbol{v}_{t-1}\}$. This set of variables enables construction of regimes that discriminate well the temporal variability of the process $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$. Let denote $\mathrm{C}_{Diff(u,v)}$ this second local clustering.

5 Analysis of the proposed clusterings

The proposed clusterings are compared through various analyses. We seek a clustering that is physically meaningful and appropriate in terms of conditional autoregressive models. For a proper comparison, for all clusterings, we decide to order regimes from the more persistent to the less persistent. This is done according to the determinant of the matrice $\boldsymbol{\Sigma}^{(i)}$.

5.1 First visual comparison

Sequences of regimes from the proposed clusterings are shown in Figure 5. The top panel shows that C_{Z500} has very persistent regimes. This result is expected because it describes the alternation between the preferred states of the large-scale atmospheric dynamics, whose typical time scale is a few days. One can see that the less volatile wind conditions are associated with the BL and AR phases, whereas the most variable wind conditions occur during the two NAO phases; see Figure 10. The three bottom panels correspond to local clusterings. For all of them, the first regime is associated with the less volatile conditions with weakest intensity, whereas the second and third regimes are generally associated with moderate and high intensity of wind. However, the behavior of the regime-switching differs from one clustering to another, probably because of the different choice of descriptors ($\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$ vs. $\{\boldsymbol{u}_t - \boldsymbol{u}_{t-1}, \boldsymbol{v}_t - \boldsymbol{v}_{t-1}\}$) and/or methods (observed vs. latent) used in the clustering. The bottom panel of Figure 2 shows that the second regime is a precursor

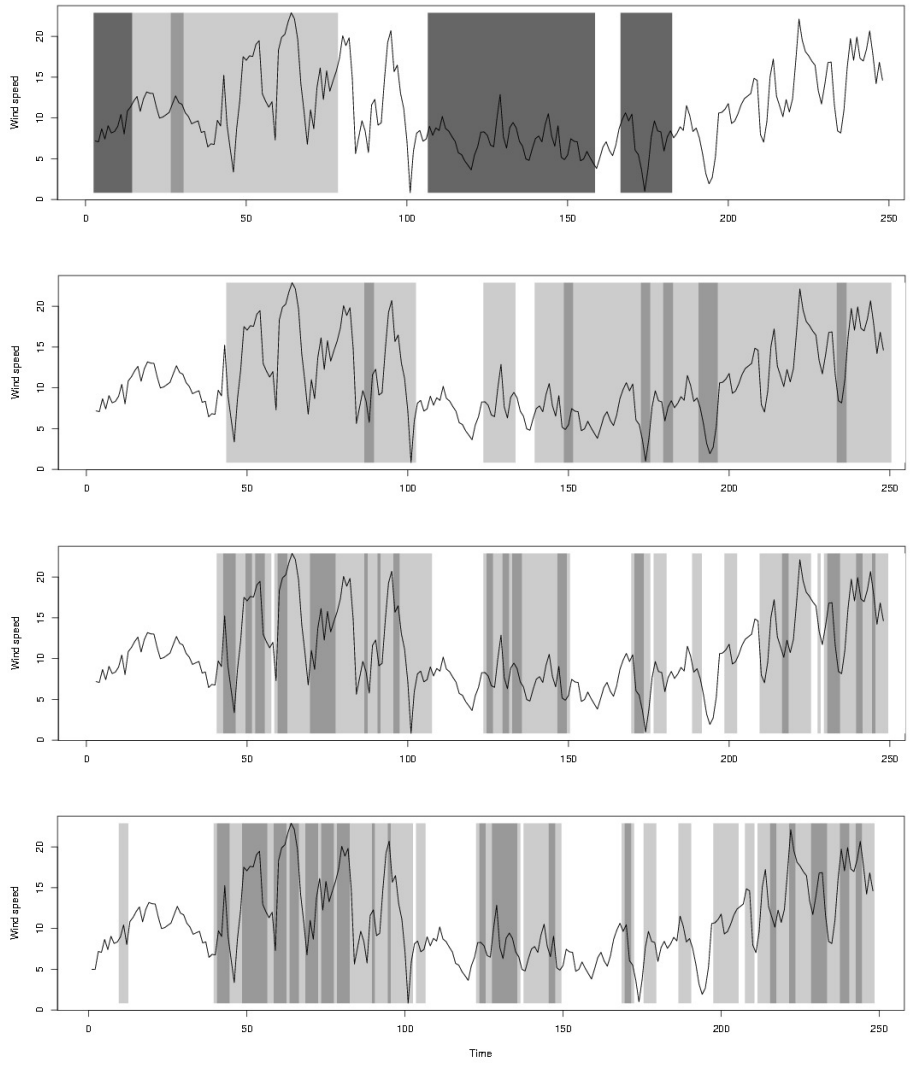


Figure 5. Time series of wind speed in January 2012 and a priori regimes extracted from the proposed methods above. The darker is the grey; the smaller is the determinant of $\mathbf{\Sigma}^{(i)}$. From top to bottom: C_{Z500}, $\mathrm{C}_{EOF(u,v)}$, $\mathrm{C}_{Diff(u,v)}$, and regimes from the fitting of the H-MS-VAR model.

to the third one (which is confirmed by the transition probabilities between regimes) and that this second regime is most of the time associated with rises in wind speed intensity.

In Figure 6, the average fields corresponding to each regime of the four clusterings are plotted. The top row highlights the difficulty of discriminating local wind features when using regimes defined from a large-scale circulation variable. While the AR and NAO+ regimes of C_{Z500} are associated with strong local wind signatures (as described in Subsection 4.1), the BL and NAO− regimes have a weaker discriminatory power on the local wind data. This issue was also observed in (Najac, 2008).

Since different descriptors are used, $C_{Diff(u,v)}$ and $C_{EOF(u,v)}$ lead to very different results. $C_{EOF(u,v)}$ leads to the most physically consistent regimes: a northeasterly regime, a northwesterly one, and a southwesterly one, which are flows corresponding to several of the large-scale weather regimes. The last two regimes are associated with stronger intensities. From the derivation of this clustering, one naturally finds regimes that correspond to the main mean patterns of variability of the fields.

The regimes of $C_{Diff(u,v)}$ have less persistence, which complicates their meteorological interpretation. The first regime corresponds to periods of weak wind intensities. The last two regimes are southwesterly regimes with different intensity from one to the other. The averaged fields of the regimes extracted from H-MS-VAR are similar to the ones of $C_{Diff(u,v)}$ despite some punctual discrepancies in their time series (Figure 5). The first regime of these two clusterings seems associated with blocking situations.

To compare the associations between the different classifications, a multiple correspondance analysis is made between the four categorical variables that represent each classification. This analysis can be viewed as an analog of a Principal Component Analysis for categorical variables where the associations between the variables are measured with the Chi-squared distance. The regimes of each classification are projected on the two first components and displayed in Figure 7. These two axes enable to account for 44% of the variance, which is not low for such an analysis. The other axes are not considered because they do not bring enough useful information. Notice that this analysis does not account for the temporal dependence in each classification. The overall structure tends to associate the three classifications $C_{EOF(u,v)}$, $C_{Diff(u,v)}$ and H-MS-VAR, except for the third regime of $C_{EOF(u,v)}$. The classifications $C_{Diff(u,v)}$ and H-MS-VAR are very close in this projection, which means that their regimes mainly occur at the same time and this coincides with the Figure 5. The first axe opposes time persistent regimes to less persistent ones. The regime BL is close to the regimes R1 of $C_{EOF(u,v)}$, $C_{Diff(u,v)}$ and H-MS-VAR, this is also seen in Table 3 and is in agreement with the average fields of these regimes displayed on Figure 6. The second axe opposes the regimes R2 of H-MS-VAR and $C_{Diff(u,v)}$ from the regimes R3, which is also an opposition from persistent to less persistent regimes. Most of these similarities between the regimes are also seen on Table 2 through the logarithm of the covariance of the innovations and the percentage of time spent in each regime. The regime AR from C_{Z500} seems more difficult to associate with other regimes. The regime R3 from $C_{EOF(u,v)}$ is associated to the weather regime NAO+, which coincides with Table 3 and Figure 6.

5.2 Quantitative analyzing

Quantitative criteria are considered in order to complete this analysis. The optimal value of the complete log-likelihood of the model is generally a good measure of the statistical relevance of a model. The complete log-likelihood, given in (3), evaluated at the maximum likelihood estimator of

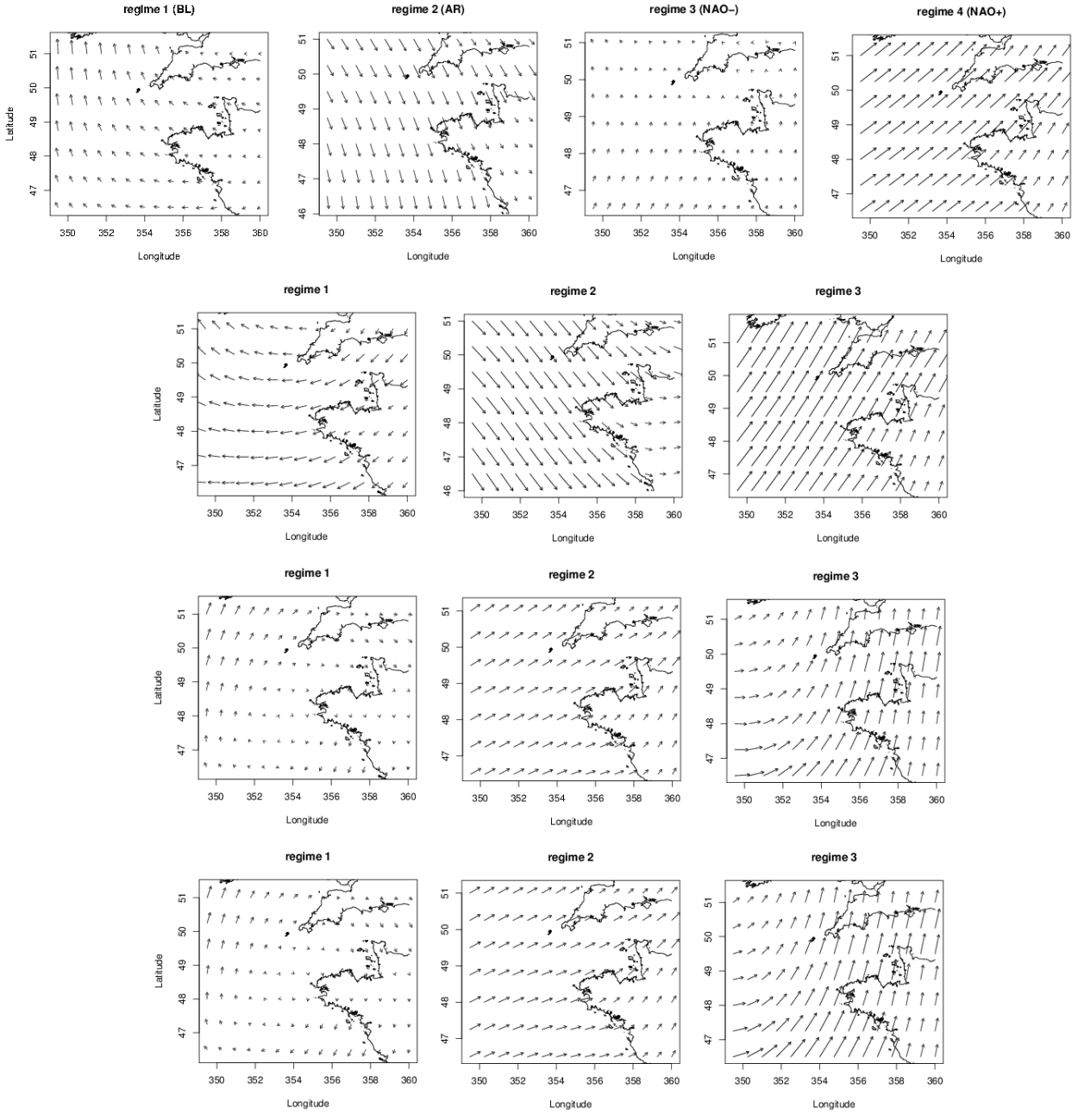


Figure 6. Average fields of $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$ in each regime of the clusterings, from top to bottom: C_{Z500}, $\mathrm{C}_{EOF(u,v)}$, $\mathrm{C}_{Diff(u,v)}$ and from the fitting of H-MS-VAR on the set of 5 locations.

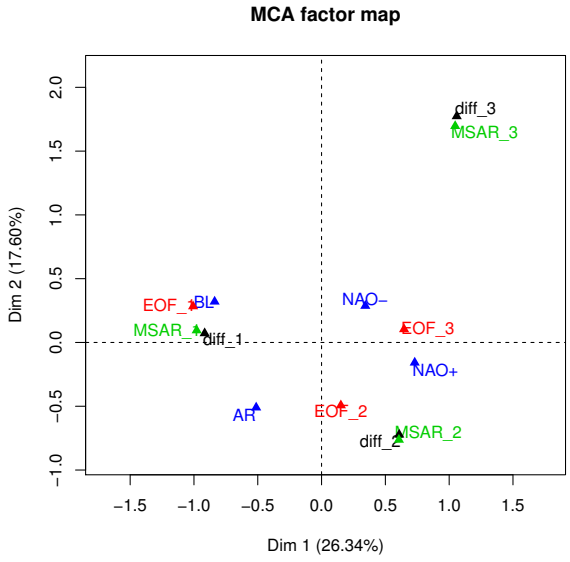


Figure 7. First plan of the multiple correspondance analysis made on the four classifications. Each regime of the four classification are depicted.

$\hat{\theta}$, is written in the case of observed regime-switching as the sum of the two following terms:

$$\log(\mathcal{L}(\hat{\theta}^{(\boldsymbol{Y})}; \boldsymbol{y}_1^T | s_1^T)) = -\frac{Td\log(2\pi)}{2} - \frac{Td}{2} - \sum_{i=1}^{M} n_i \log(\det(\hat{\boldsymbol{\Sigma}}^{(i)}))$$

and

$$\log(\mathcal{L}(\hat{\theta}^{(S)}; s_1, ..., s_T)) = \sum_{i,j=1}^{M} n_{i,j} \log\left(\frac{n_{i,j}}{n_{i,\cdot}}\right).$$

Note that the first term is a function of the total time spent in each regime and the associated determinant of covariance matrix of innovation (notice that the one-step-ahead error of the forecast is linked to this quantity). The longer the time spent in a regime with a weak determinant of covariance of innovation, the greater the log-likelihood (see Table 2). The maximal log-likelihood of $\theta^{(S)}$ is equal to the opposite of the conditional entropy of S_t given S_{t-1}. The conditional entropy is classically used as a quality measure of clustering. In prediction, the weaker the entropy, the stronger the predictability of S_t given S_{t-1}. More generally one tends to minimize this measure. Because of the range of values of the log-likelihood of $\theta^{(\boldsymbol{Y})}$, the value of that of $\theta^{(S)}$ has a low contribution to the complete log-likelihood. If the complete log-likelihood is used to select models, the persistence of the Markov chain has a low impact. BIC indexes are also given in Table 2, where $\mathrm{BIC} = -2\log \mathrm{L} + \mathrm{N}_p \log(\mathrm{N}_{obs})$ with L the likelihood of the model, N_p the number of parameters and N_{obs} the number of observations. The BIC index enables one to consider a compromise between

Table 2. N_p the number of parameters. Values are computed from models fitted on $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$ at the 5 locations (1,6,10,13,18).

Model	BIC	log-$\mathcal{L}$ of S	log-$\mathcal{L}$ of $\boldsymbol{Y}$	N_p	$\log(\det(\boldsymbol{\Sigma}^{(i)}))$				% of Time Spent			
					R1	R2	R3	R4	R1	R2	R3	R4
Unconditional VAR	542640	-	-269825	265	36.4	-	-	-	-	-	-	-
AP-MS-VAR$_{C_{Z500}}$	542730	-1510	-263808	1072	29.8	30.3	39	38.1	0.27	0.18	0.2	0.34
AP-MS-VAR$_{C_{EOF(u,v)}}$	545730	-2331	-266015	801	28.9	33.3	38.9	-	0.31	0.42	0.27	-
AP-MS-VAR$_{C_{Diff(u,v)}}$	520759	-4762	-251099	801	20.2	34.1	48.1	-	0.44	0.41	0.15	-
H-MS-VAR	459458	-	-229616	801	18.4	32.1	48.4	-	0.43	0.41	0.16	-

Table 3. Joint probability of occurrence of the three local regimes identified by the proposed models in rows and the four large-scale regimes in columns

	$C_{EOF(u,v)}$					$C_{Diff(u,v)}$					H-MS-VAR				
	BL	AR	NAO−	NAO+	Total	BL	AR	NAO−	NAO+	Total	BL	AR	NAO−	NAO+	Total
R1	0.17	0.06	0.08	0.01	0.32	0.15	0.10	0.07	0.13	0.45	0.13	0.09	0.07	0.14	0.43
R2	0.04	0.10	0.05	0.08	0.27	0.09	0.06	0.09	0.16	0.40	0.10	0.06	0.09	0.15	0.41
R3	0.07	0.02	0.07	0.26	0.42	0.03	0.02	0.04	0.06	0.15	0.04	0.02	0.05	0.06	0.16
Total	0.28	0.18	0.20	0.35	1	0.27	0.18	0.20	0.35	1	0.27	0.17	0.21	0.35	1

a model with a high likelihood and its parsimony. Notice that one should not compare BIC indexes of a priori and of latent regime-switching models. However the BIC indexes of these two classes of models can be compared with that of the unconditional VAR model, since it is a particular case.

The clustering $C_{Diff(u,v)}$ provides the greatest value of complete log-likelihood. The lower value of log-likelihood of S, with shorter persistence in the different regimes compared with the other models, is compensated by a larger value of log-likelihood of Y and thus a longer time spent in regimes with low variances of innovation. The three proposed AP-MS-VAR models lead to a satisfying description of the marginal and joint distributions and space-time covariances (not shown). The model AP-MS-VAR$_{C_{Diff(u,v)}}$, which exhibits the best likelihood, performs the most accurately among the AP-MS-VAR models to reproduce the moving average and moving variance processes; see Section 6. Besides in terms of BIC indexes, the smallest value among the AP-MS-VAR models is that of AP-MS-VAR$_{C_{Diff(u,v)}}$ and it is also greater than that of the VAR model. In the following, the VAR model with shifts defined by $C_{Diff(u,v)}$ is kept for further comparisons with the H-MS-VAR model in simulation; see Section 6. We choose this model although it is not the most physically meaningful because it leads to better results according to our criterion.

5.3 Link between large-scale weather regimes and local ones

In this section we quantitatively compare the large-scale regimes described by C_{Z500} with the local ones derived from the hidden MS-VAR. To this end, we compute the joint probability of occurrence of large-scale regimes (C_{Z500}) and local regimes (successively $C_{EOF(u,v)}$, $C_{Diff(u,v)}$ and H-MS-VAR, Table 3).

For the three clusterings, the local regimes seem to appear in preferential large-scale weather regimes. The strongest link with C_{Z500} is found for $C_{EOF(u,v)}$: the first regime coincides mainly with BL, the second one with AR, and the third one with NAO+. These results are not surprising because regimes of $C_{EOF(u,v)}$ are also easier to interpret physically. However, the association is not systematic: for instance, the second regime is observed not only during AR conditions but also during NAO+ conditions. Note that NAO− conditions split rather equiprobably among the three local regimes.

The regimes of H-MS-VAR and of $C_{Diff(u,v)}$ are more difficult to link with large-scale regimes. The fact that they are less persistent than the $C_{EOF(u,v)}$ ones may explain why their joint occurrences with C_{Z500} are weaker. As previously said, H-MS-VAR regimes are driven mainly by the conditional autoregressive model in the sense of the likelihood, which results in a more difficult physical interpretation. Some links can nevertheless be made: for both H-MS-VAR and $C_{Diff(u,v)}$, the second regime coincides mainly with NAO+, and to a lesser extent the first regime is connected to BL.

6 Comparison in simulation of the multisite wind models

In this section, we compare models VAR(2), AP-MS-VAR$_{C_{Diff(u,v)}}$ and H-MS-VAR in terms of reproducing the various scales of the spatiotemporal wind variability. We focus on the alternation between periods with different temporal variability of wind conditions, and we highlight the benefit of using appropriate regime-switching in reproducing such an alternation. $N = 100$ sequences of the length of the data are generated with the fitted models and several statistics are computed on these data.

First, marginal statistics at the central site 10 are investigated (see Figure 8). Comparing Figures 1 and 8, one can notice that the distribution of $\{u_t\}$ is well reproduced by the model H-MS-VAR, while the $\{v_t\}$ one is less accurately described. Results in (Ailliot, Bessac, Monbet, and Pene, 2015) are slightly more satisfying because of non homogeneous transitions between regimes. The description of this distribution by AP-MS-VAR$_{C_{Diff(u,v)}}$ is also satisfying and not shown here. Concerning the temporal dependence, the regime-switching models are the most able to accurately reproduce the autocorrelation functions of both $\{u_t\}$ and $\{v_t\}$. All the models tend to behave similarly in reproducing the correlation of $\{u_t\}$. However, the VAR model tends to underestimate the dependence of $\{v_t\}$ between 2 and 5 days, and the regime-switching models improve the description of this dependence.

The space-time correlation function of the multivariate process $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$ and its simulated replicates reveals that both models reproduce satisfyingly the general shape of this function and especially the non separable and anisotropic patterns; see Figure 9. The non separability is reflected in the asymmetry around the vertical axis at lag 0 is captured by the proposed models.

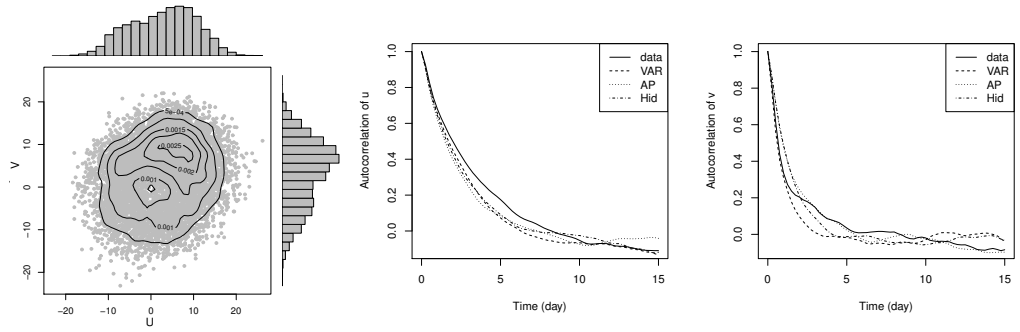


Figure 8. Left: joint and marginal distribution of simulated data at site 10 from the model H-MS-VAR. Central and right panels: autocorrelation functions of $\{u_t\}$ and $\{v_t\}$ at site 10 for the reference data, simulated data from the VAR(2), AP-MS-VAR$_{C_{Diff(u,v)}}$ and H-MS-VAR models.

To study patterns at an instantaneous time scale, we focus on the ability of the models to reproduce the alternation of temporal variability. Indeed the alternation of different weather states induces an alternation in the intensity and temporal variability of wind. In Figure 10, the moving standard deviation of wind speed around its moving mean at the central site 10 is depicted as a function of its moving mean. Observations reveal a higher variability when the intensity is high, although a high variability may also be associated with weaker values when the moving window overlaps the transition time. Models with regime-switching enable the reproduction of more temporal variability associated with moderate and high intensity of wind, which is not captured by an unconditional VAR model. For instance, the regime-switching models reproduce high variability around 5 and 10 $m.s^{-1}$ which corresponds to transitions between weather states. This is ensured by the alternation, driven by a Markov chain, of periods associated with different parameters of the conditional model. Similar diagnostics than in Figure 4 indicate that the distributions of the moving standard deviation and the moving mean within each simulated regime of the C$_{Diff(u,v)}$ and of H-MS-VAR are clearly distinct from one regime to the other, which indicates characteristic behaviors of these two simulated processes within each regime (not shown). Moreover, the behavior in each simulated regime is close to the observed one.

7 Discussions and perspectives

In Section 3, we compare site-specific regimes to common regional regimes. We conclude according to mainly qualitative criteria that for this dataset the use of a regime common to all locations is reasonable. To go one step further, one would settle some likelihood-ratio test, to quantify more precisely to which extent the assumption of a regional regime against a site-specific regime is acceptable.

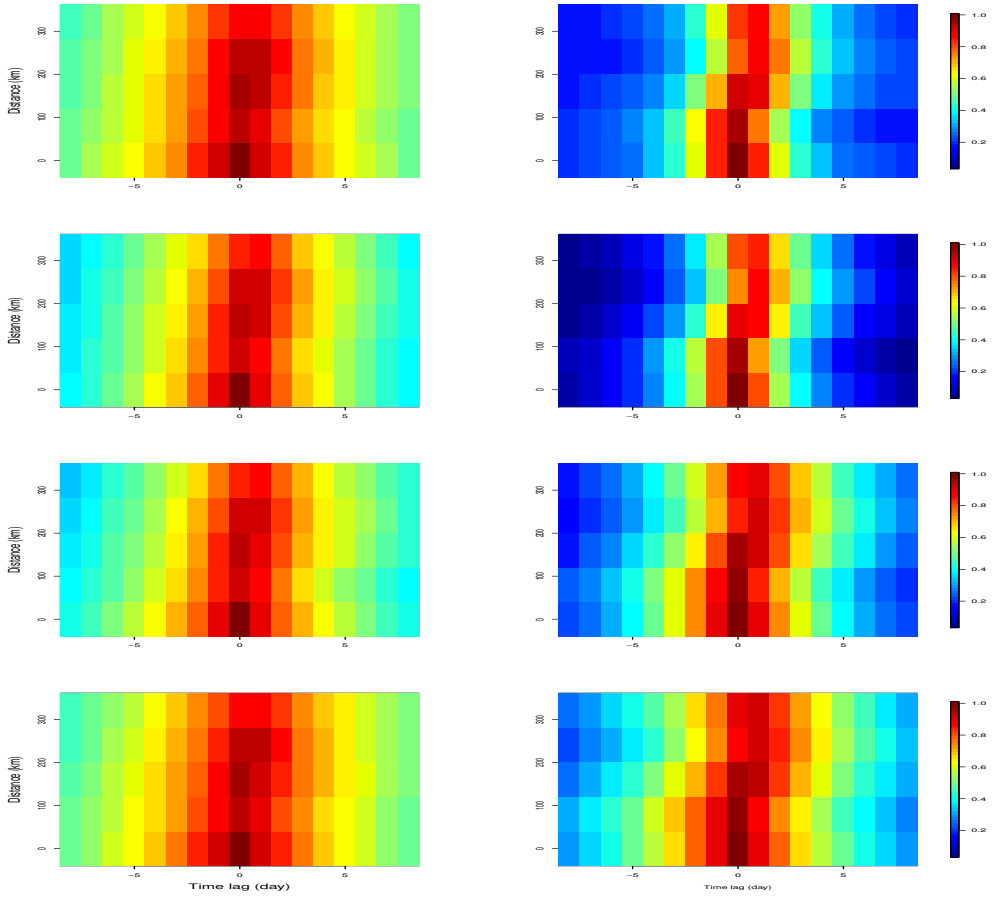


Figure 9. Left: correlation of between $\{u_t\}$ at site 1 and $\{u_t\}$ at the other locations (sorted according increasing distance) at various time-lag. Right: similar quantities for $\{v_t\}$. From top panel to the bottom one: data, simulation from VAR(2), from AP-MS-VAR$_{C_{Diff(u,v)}}$, and from H-MS-VAR.

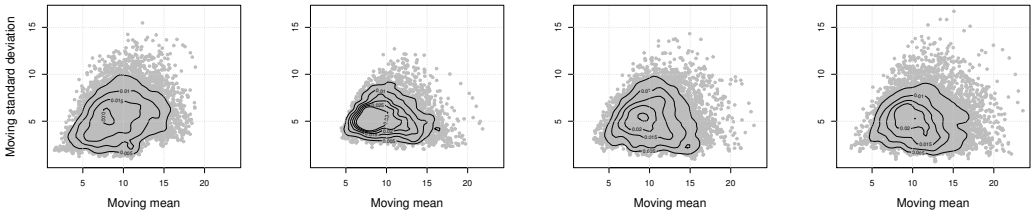


Figure 10. Moving standard deviation against of the value $\{U_t\}$ against its moving mean at location 10. From left to right: data, simulation from the VAR(2), AP-MS-VAR$_{C_{Diff(u,v)}}$, and H-MS-VAR

In this paper we have introduced observed and latent regime-switching framework, and we have showed that both types of regime-switching models have various advantages. Models with observed switchings may account for relevant regimes that correspond to characteristic meteorological conditions in Europe. The choice of the clustering method and of the descriptors of the data is crucial, as discussed in Subsection 4.2 where a k-means clustering led to irrelevant regimes in terms of estimation of the associated conditional model.

The hidden regime-switching framework seems to overcome this insufficiency by providing regimes that are driven by the conditional distribution and therefore adapted to the estimation. When considering hidden regime-switching models, however, the estimation procedure may become challenging when sophisticated marginal models are considered. The extracted regimes are driven mainly by the local data and the proposed conditional distribution, and consequently they might have less physical interpretation than do regimes derived from other clusterings. Nevertheless, in this study we saw that for the proposed model and studied dataset, the associated regimes were not physically inconsistent. Moreover, the use of hidden regime-switching models saves efforts in choosing an appropriate observed a priori clustering.

Concerning the proposed observed regime-switching models, there seems to be a compromise between physically interpretable regimes and a good description of the conditional model by a VAR, as highlighted in Section 4 when comparing AP-MS-VAR$_{\mathrm{C}_{Diff(u,v)}}$ and AP-MS-VAR$_{\mathrm{C}_{EOF(u,v)}}$ models. Indeed we have chosen AP-MS-VAR$_{\mathrm{C}_{Diff(u,v)}}$ because it provides the best BIC index despite the fact that $\mathrm{C}_{Diff(u,v)}$ has less physical interpretation. This highlights the difficulty in finding relevant regimes that are adapted to the description of the data by conditional vector autoregressive models. The proposed hidden regime-switching model seems to respond to this compromise in providing more interpretable regimes than the ones of $\mathrm{C}_{Diff(u,v)}$ and similar description of temporal patterns. The improvement of BIC from the AP-MS-VAR$_{\mathrm{C}_{Diff(u,v)}}$ with respect to the unconditional VAR is 4% whereas the improvement from the H-MS-VAR is15.3%.

Future work may involve investigating reduced parameterizations of the autoregressive coefficients and of the matrices of covariance of innovations, thus helping to adapt the model to a larger dataset. Indeed the number of parameters is already high with the small dataset under consideration, and attempts to use parametric shapes for parameters reveal that a huge effort will be needed to extract consistent results. Furthermore, when looking at the autoregressive matrices, one sees generally privileged predictors according to the regimes, a situation that motivates the use of constraint matrices in each regime.

References

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

Ailliot, P., Monbet, V., and Prevosto, M.: An autoregressive model with time-varying coefficients for wind fields, Environmetrics, 17, 107–117, 2006.

Ailliot, P., Thompson, C., and Thomson, P.: 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, 405–426, 2009.

Ailliot, P., Bessac, J., Monbet, V., and Pene, F.: Non-homogeneous hidden Markov-switching models for wind time series, Journal of Statistical Planning and Inference, 160, 75–88, 2015.

Bardossy, A. and Plate, E. J.: Space-time model for daily rainfall using atmospheric circulation patterns, Water Resources Research, 28, 1247–1259, 1992.

Bessac, J., Ailliot, P., and Monbet, V.: Gaussian linear state-space model for wind fields in the North-East Atlantic, Environmetrics, 26, 29–38, 2015.

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

Cassou, C.: Intraseasonal interaction between the Madden–Julian oscillation and the North Atlantic oscillation, Nature, 455, 523–527, 2008.

Cattiaux, J., Douville, H., and Peings, Y.: European temperatures in CMIP5: origins of present-day biases and future uncertainties, Climate Dynamics, 41, 2889–2907, doi:10.1007/s00382-013-1731-y, 2013.

Dempster, A. P., M., L. N., and Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 39, 1–38, 1977.

Durand, J.-B.: Modèles à structure cachée: inférence, estimation, sélection de modèles et applications, Ph.D. thesis, Université Joseph-Fourier-Grenoble I, 2003.

Flecher, C., Naveau, P., Allard, D., and Brisson, N.: A stochastic daily weather generator for skewed data, Water Resources Research, 46, W07 519, 2010.

Fuentes, M., Chen, L., Davis, J. M., and Lackmann, G. M.: Modeling and predicting complex space–time structures and patterns of coastal wind fields, Environmetrics, 16, 449–464, 2005.

Gneiting, T., Larson, K., Westrick, K., Genton, M. G., and Aldrich, E.: Calibrated probabilistic forecasting at the stateline wind energy center: The regime-switching space–time method, Journal of the American Statistical Association, 101, 968–979, 2006.

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

Hamilton, J. D.: Analysis of time series subject to changes in regime, Journal of Econometrics, 45, 39–70, 1990.

Haslett, J. and Raftery, A. E.: Space-time modelling with long-memory dependence: Assessing Ireland's wind power resource, Applied Statistics, pp. 1–50, 1989.

Hering, A. S. and Genton, M. G.: Powering up with space-time wind forecasting, Journal of the American Statistical Association, 105, 92–104, 2010.

Hering, A. S., Kazor, K., and Kleiber, W.: A Markov-switching vector autoregressive stochastic wind generator for multiple spatial and temporal scales, Resources, 4, 70–92, 2015.

Hofmann, M. and Sperstad, I. B.: NOWIcob–A tool for reducing the maintenance costs of offshore wind farms, Energy Procedia, 35, 177–186, 2013.

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

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

Khalili, M., Leconte, R., and Brissette, F.: Stochastic multisite generation of daily precipitation data using spatial autocorrelation, Journal of Hydrometeorology, 8, 396–412, 2007.

Kleiber, W., Katz, R. W., and Rajagopalan, B.: Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes, Water Resources Research, 48, n/a–n/a, doi:10.1029/2011WR011105, http://dx.doi.org/10.1029/2011WR011105, 2012.

Michelangeli, P. A., Vautard, R., and Legras, B.: Weather regimes: recurrence and quasi stationarity, Journal of the Atmospheric Sciences, 52, 1237–1256, 1995.

Najac, J.: Impacts du changement climatique sur le potentiel éolien en France: une étude de régionalisation, Ph.D. thesis, Université Paul Sabatier-Toulouse III, 2008.

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

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

Thompson, C. S., Thomson, P. J., and Zheng, X.: Fitting a multisite daily rainfall model to New Zealand data, Journal of Hydrology, 340, 25–39, 2007.

Vautard, R.: Multiple weather regimes over the North Atlantic: Analysis of precursors and successors, Monthly Weather Review, 118, 2056–2081, 1990.

Vrac, M., Stein, M., and Hayhoe, K.: Statistical downscaling of precipitation through nonhomogeneous stochastic weather typing, Climate Research, 34, 169, 2007.

Wikle, C. K., Milliff, R. F., Nychka, D., and Berliner, L. M.: Spatiotemporal hierarchical Bayesian modeling tropical ocean surface winds, Journal of the American Statistical Association, 96, 382–397, 2001.

Wilks, D. S.: Multisite generalization of a daily stochastic precipitation generation model, Journal of Hydrology, 210, 178–191, 1998.

Wilson, L. L., Lettenmaier, D. P., and Skyllingstad, E.: A hierarchical stochastic model of large-scale atmospheric circulation patterns and multiple station daily precipitation, Journal of Geophysical Research: Atmospheres (1984–2012), 97, 2791–2809, 1992.

Wu, C. F. J.: On the convergence properties of the EM algorithm, Annals of Statistics, 11, 95–103, 1983.

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

Zucchini, W. and MacDonald, I.: Hidden Markov models for time series: An introduction using R, no. 110 in Monographs on statistics and applied probability, CRC Press, 2009.