HAL
open science

On the construction of stochastic generators of wind conditions offshore Brittany

Julie Bessac

To cite this version:

Julie Bessac. On the construction of stochastic generators of wind conditions offshore Brittany. General Mathematics [math.GM]. Université de Rennes, 2014. English. NNT : 2014REN1S067 . tel-01079520

HAL Id: tel-01079520

https://theses.hal.science/tel-01079520v1

Submitted on 4 Nov 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.

N° d'ordre : ANNÉE 2014

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Mathématiques et applications

Ecole doctorale MATISSE

présentée par

Julie Bessac

préparée à l'UMR 6625 CNRS-IRMAR
Institut de recherche Mathématique de Rennes
U.F.R. de Mathématiques

Sur la construction de générateurs aléatoires de conditions de vent au large de la Bretagne

Thèse soutenue à Rennes le 20 octobre 2014

devant le jury composé de :

Pierre AILLIOT
Maître de conférences, Université de Brest / Directeur de thèse

Denis ALLARD
Directeur de recherche INRA / Rapporteur

Laurent MEVEL
Chargé de recherche INRIA / Examinateur

Valérie MONBET
Professeur, Université de Rennes 1 / Directeur de thèse

Pierre PINSON
Professeur, Technical University of Denmark / Examinateur

Igor RYCHLIK
Professeur, Chalmers University of Technology / Rapporteur

Remerciements

Je remercie en premier lieu mes directeurs de thèse Pierre Ailliot et Valérie Monbet, de m'avoir fait découvrir le monde de la modélisation et de la recherche. Merci pour cette aventure statistique et humaine, pour la considération que vous m'avez portée dès le début, pour votre engagement dans ce travail et pour le partage de vos connaissances tant scientifiques qu'humaines.

Je tiens à remercier Denis Allard et Igor Rychlik d'avoir accepté d'être rapporteur de ma thèse et pour l'intérêt qu'ils ont porté à mon travail. Je remercie Pierre Pinson et Laurent Mevel d'avoir accepté de faire partie de mon jury.

Je remercie les professeurs du Magistère, les professeurs de l'ENS Ker Lann et les professeurs de l'Université, que j'ai eus au cours de ma scolarité à Rennes, pour cette formation mathématique et humaine que vous dispensez avec passion et intérêt. Je remercie en particulier Arnaud Debussche pour son encadrement et Mihai Gradinaru pour ses précieux conseils, son écoute sincère et son soutien tout au long de mon parcours à Rennes.

Je remercie François Coquet et Magalie Fromont, avec qui j'ai découvert la statistique, de la considération qu'ils m'ont toujours portée.

Ce fut un très grand plaisir d'enseigner à l'ENSAI, je remercie les enseignants avec qui j'ai eu le plaisir de travailler François Coquet, Myriam Vimond et Salima El Kolei ; Emilie Chautru et Marie-Anne Vibet avec qui nous avons discuté tests statistiques mais pas que.

Peter and Erris Thomson, I thank you very much for welcoming me so warmly in New-Zealand. Peter, working with you was very enriching, motivating and a great honor.

Je remercie les chercheurs que j'ai rencontrés lors de conférences : Philippe Naveau, Liliane Bel, Julie Carreau pour leur accueil et leurs encouragements lors de mes premiers pas en conférences, ainsi que Matthieu Vrac pour ses conseils.

Je remercie chaleureusement Aurélien Ribes et Julien Cattiaux pour leur dynamisme et leur sympathie à toute épreuve, j'espère que nous aurons encore l'occasion de travailler ensemble.

Je remercie les chercheurs de l'IRMAR pour un sourire ou un bonjour, je tiens à remercier Monique Dauge pour ses conseils francs et avisés.

Je remercie très chaleureusement les membres du personnel de l'IRMAR

et de l'UFR pour leur disponibilité, leur efficacité et leur gentillesse : Chantal, Hélène, Marie-Aude, Emmanuelle, Xhensila, Patrick, Marie-Annick ; merci Olivier pour tous vos services. Je remercie également Élodie Cottrel et Anne-Joëlle Chauvin.

Doctorants statisticiens rennais : Samuel, Gaspar, Cyril, Emeline ; copains de conférence : Marc Bourotte, Romain Chailan et Aurélien Bechler ; les matheux du plateau de Ker Lann : Thibaut, Guillaume et Quentin, merci pour votre bonne humeur, c'était et c'est toujours un plaisir de vous croiser.

Marie et Tiphaine, merci les filles, c'était vraiment un plaisir de partager le bureau avec vous et les soirées filles. Thank you Katharina for your support and our endless discussions.

Merci aux amis et copains rencontrés à l'IRMAR tout au long de cette thèse, je pense aux doctorants du bureau 434, post-docs ou autre PU aux biceps gonflables : Daminou, Elise, Kodjo, Alexandre, Maher, Christophe Tran, Christophe Ritzenthaler, Jeroen, Marianna.

Les amis et copains du magistère Aurore et Bob, Pierre et Eugénie, Claire-Soizic, Christophe et Angélina, merci pour les goûters post-BU du samedi en prépa agreg, les week-ends, rando-vélo et tout le reste !

Merci aux grimpeurs de m'avoir permis d'oublier la troisième année de thèse quelques soirs par semaine : Marine (dite Profitor) et Pierre (Le Guignolo).

Pierre-Antoine, je remercie tes ligaments croisés de nous faire passer de supers vacances au ski. Merci Charon pour cette description de l'anatomie de Graou, merci Jimmy et Sandra pour votre sincère soutien et votre chaleur humaine.

Merci aux copines que j'ai rencontrées un peu trop tard à Rennes : Carole, médecin en mal de statisticien et Leslie.

Elo, la super-girl, merci pour tes encouragements et ta joie de vivre.

Je remercie Rominou, mon colocataire et ami, qui a supporté (non sans peine j'imagine) ma troisième année de thèse, ma phobie de l'inactivité et mes indécisions.

Je remercie mes parents pour leur soutien à toute épreuve et pour tout ce que vous avez fait pour moi, merci à ma famille pour leurs encouragements.

Nicolas, supra-méta-merci pour ton très précieux soutien, ton écoute attentive tout au long de cette thèse, tes conseils mais surtout pour ta passion débordante pour la vie.

Contents

Résumé **1**

1 Introduction **11**
1.1 Stochastic weather generators 12
1.2 Meteorological data 13
1.2.1 Various sources of data 13
1.2.2 Considered wind data and associated issues 14
1.3 Hidden Markov and state-space models 20
1.3.1 Basic properties 20
1.3.2 Inference 22
1.4 Modeling space-time dependence 23
1.4.1 Temporal dependence at a single site 23
1.4.2 Including multi-site interactions 24
1.4.3 Accounting for regime-switching 25
1.5 Interaction between wind direction and speed 27
1.5.1 Modeling wind direction time series 27
1.5.2 Modeling polar and Cartesian coordinates of wind . . . 28
1.6 Plan of the thesis 30

2 A multi-site model for wind speed **31**
2.1 Introduction 31
2.2 The wind dataset 33
2.3 A linear Gaussian state-space model 36
2.3.1 Model 36
2.3.2 Second-order structure and identifiability 37
2.3.3 Parameter estimation 38
2.3.4 Properties of the estimates 40
2.4 Results 40
2.4.1 Interpretability 40
2.4.2 Realism of simulated sequences 41
2.5 Some improvements of the model 43
2.5.1 Parameterization of Γ 43
2.5.2 Parameterization of Λ 46

2.6 General discussion 49
2.7 Proof of proposition 1 49
2.8 Maximum Likelihood Estimation 51
2.8.1 Kalman recursions 52
2.8.2 EM algorithm 53
2.9 Prediction as a validation tool 54

3 Single-site models for wind time series 57
3.1 Introduction 57
3.2 Models 59
3.2.1 Non-homogeneous Markov-switching autoregressive models 59
3.2.2 Non-homogeneous Markov model for the weather type 61
3.2.3 Modeling the Cartesian coordinates conditionally to the weather type 62
3.2.4 Modeling the wind direction conditionally to the weather type 63
3.2.5 A conditional model for the wind speed given the wind direction 64
3.3 Parameter estimation 66
3.3.1 Numerical computation of the MLE 66
3.3.2 Ergodicity of the models and asymptotic properties of the MLE 67
3.4 Numerical results and model comparison 71
3.4.1 Data 71
3.4.2 Model selection 72
3.4.3 Regimes can be interpreted as weather types 73
3.4.4 Marginal distributions 76
3.4.5 Dependence structure 77
3.5 Conclusion 78

4 Multi-site models for (u, v)-time series 81
4.1 Introduction and general context 82
4.1.1 Introduction 82
4.1.2 Wind data 83
4.2 Markov-Switching AutoRegressive models 85
4.2.1 The models 85
4.2.2 Estimation by maximum likelihood 85
4.3 From a single-site to a multi-site hidden MS-AR model 88
4.4 Observed regime-switching models 92
4.4.1 Derivation of observed regimes from extra-variables . . 92
4.4.2 Derivation of observed regimes from the local variables 93
4.4.3 Comparison and selection 93

4.5 Comparison of the multi-site wind models 97
4.5.1 Regimes extracted from hidden MS-VAR model 98
4.5.2 Link between large-scale weather regimes and the other regimes 98
4.5.3 Comparison of the MS-VAR models 99
4.6 Discussions and perspectives 100

5 Discussion **105**
5.1 Comparison of both multi-site models in simulation 105
5.2 General discussions and perspectives 108

Résumé

Nous nous proposons dans ce travail de construire des modèles stochastiques permettant de reproduire les propriétés statistiques de données spatio-temporelles de vent. Ces modèles peuvent être utilisés comme générateurs aléatoires de séquences artificielles de conditions de vent. Les séries de variables météorologiques observées couvrent des intervalles de temps généralement trop courts ou contiennent trop de valeurs manquantes pour estimer de manière fiable des probabilités d'évènements complexes. Un des objectifs de ces générateurs stochastiques de conditions météorologiques est de simuler un nombre illimité de séquences artificielles aussi longues que souhaitées. Ces séquences peuvent être utilisées dans des études d'impact mettant en jeu des variables météorologiques (voir par exemple (Skidmore and Tatarko, 1990; Hofmann and Sperstad, 2013)). Les générateurs aléatoires permettent également la simulation conditionnelle de données manquantes (Yang et al., 2005) ou la construction des scénarios de changements climatiques à échelle locale (Semenov and Barrow, 1997). La majorité des générateurs proposés tendent à reproduire au mieux certaines des propriétés statistiques observées sur les données servant à calibrer le modèle, comme par exemple la distribution en probabilité ou la fonction d'autocorrélation des variables étudiées. Par la suite, nous considérerons entre autres ces statistiques pour valider les modèles suggérés.

En introduction de cette thèse, nous proposons une description des données étudiées et des problématiques associées en vue de construire des générateurs aléatoires. Une description des modèles utilisés à savoir les modèles à espace d'états et les modèles à changement de régimes latents est donnée. Ensuite cette partie introductive est organisée selon les thématiques de modélisation mises en jeu. Pour chacune de ces thématiques, un bref état de l'art est donné ainsi que l'approche que nous proposons. Dans la première partie de ce travail, nous nous intéressons à la construction d'un modèle multi-site pour les vitesses de vent au large de la Bretagne. Nous proposons un système linéaire gaussien, nous montrons que celui-ci tend à bien reproduire la structure moyenne spatio-temporelle des données et leurs distributions marginales. Dans la suite de ce travail, nous considérons les processus bivariés en coordonnées polaires et en coordonnées cartésiennes afin de prendre en compte toute l'information contenue dans le champ de vent et de capturer l'information au-delà du comportement moyen. En premier lieu, nous nous intéressons à la

dynamique temporelle et à la distribution marginale de chaque processus bivarié en un site. La modélisation est faite à l'aide de modèles autorégressifs à changement de régimes cachés markoviens (MS-AR) avec différentes probabilités d'émission selon le couple étudié. Nous proposons ensuite une extension de ce modèle au cas multi-site pour le couple des coordonnées cartésiennes. Nous discutons la question d'un régime régional commun à toutes les stations. De plus, ce travail est l'occasion de comparer des modèles autorégressifs (AR) à changement de régimes *a priori* avec un modèle AR à changement de régimes cachés pour des données de vent. En effet ces deux types de modèles ont été largement appliqués à des données météorologiques mais sans jamais avoir été comparés.

Les modèles proposés ont été ajustés sur des données de réanalyse provenant d'ECMWF (European Center of Medium-range Weather Forecast), ces données peuvent être téléchargées gratuitement et à but scientifique à l'adresse *http://data.ecmwf.int/data/*. La zone étudiée est constituée de 18 sites situés au large de la Bretagne. Nous étudions les mois d'hiver afin d'éviter les effets saisonniers. Les modèles proposés peuvent être ajustés sur des données réelles sans aucun changement. Nous nous plaçons ici dans le cadre simplifié d'une zone rectangulaire loin des côtes afin de mettre en place la modélisation dans un contexte simple. L'objectif est ensuite de considérer une zone côtière pour laquelle les applications sont plus importantes. Cependant près des côtes, les effets locaux sont forts et difficiles à prendre en compte. Nous noterons U l'intensité du vent, Φ sa direction et u et v ses composantes zonales et méridionales.

Un modèle multi-site pour les vitesses de vent

Beaucoup de modèles pour les séries temporelles de vitesse de vent sont basés sur des modèles de type AutoRegressive Moving Average (Brown et al., 1984). Il existe peu de modèles multi-sites pour le vent et la plupart d'entre eux sont développés pour répondre à des besoins de prédiction à court terme. (Haslett and Raftery, 1989) ont proposé un modèle ARMA multi-site pour le vent en Irlande. (Rychlik and Mustedanagic, 2013) ont également construit un modèle basé sur des champs gaussiens et suggèrent une paramétrisation de la fonction de covariance spatio-temporelle.

Dans cette première partie, nous développons un modèle multi-site pour les vitesses de vent ayant pour objectif de capturer et reproduire les déplacements spatio-temporels des évènements liés aux conditions de vent. Les données météorologiques spatio-temporelles présentent fréquemment dans leur fonction de covariance une non-séparabilité des coordonnées en temps et en espace (de Luna and Genton, 2005; Finkenstädt et al., 2007). La modélisation est basée sur un système linéaire gaussien. Ces modèles ont été largement étudiés

dans (Durbin and Koopman, 2012). La forte corrélation entre les sites suggère l'utilisation d'un signal commun à chaque site contenant une majeure partie de l'information. Ce signal n'est pas directement observé et est introduit en tant que processus caché X. Nous proposons le modèle suivant :

$$(M)\left\{\begin{array}{lcl} X_{t+1} & = & \rho X_t + \sigma\epsilon_{t+1}, \\ Y_t & = & \alpha_1 X_{t+1} + \alpha_0 X_t + \alpha_{-1} X_{t-1} + \Gamma^{1/2}\eta_t \end{array}\right. \qquad \text{pour } t \geq 0,$$

$Y_t \in \mathbb{R}^K$, où K est le nombre de sites étudiés, Y représente le vent observé en chaque site, $\rho > 0$ et $\sigma > 0$. Les coefficients α_1, α_0 et α_{-1} sont des vecteurs K-dimensionnels, ϵ et η sont des bruits blancs gaussiens indépendants.

L'identifiabilité des paramètres a été étudiée via la structure d'ordre 2 du processus gaussien Y, elle est obtenue au signe près sous les hypothèses suivantes : $\frac{\rho}{1-\sigma^2}$ est fixe et les vecteurs α_1, α_0 et α_{-1} sont libres. Nous montrons également que sous ces conditions d'identifiabilité la fonction de covariance spatio-temporelle du processus multivarié Y est non-séparable.

Deux méthodes d'estimation ont été implémentées et comparées, l'une basée sur la méthode des moments généralisée et l'autre sur le maximum de vraisemblance via l'algorithme Expectation-Maximization. Chacune des méthodes présente des avantages divers tant du point de vue de l'implémentation des calculs que du point de vue des résultats. En simulation, le modèle estimé par la méthode des moments généralisée décrit plus précisement la structure temporelle à très court terme que lorsque celui-ci est estimé par maximum de vraisemblance. Tandis que ce dernier reproduit mieux la structure temporelle à plus long terme.

Différents modèles réduits ont été étudiés, notamment afin de paramétrer différentes quantités en fonction de la latitude et la longitude. Il s'avère que la paramétrisation de α_1, α_0 et α_{-1} en une fonction quadratique en latitude et longitude est simple à mettre en place et satisfaisante du point de vue des résultats en simulation. La paramétrisation de Γ avec des modèles classiques (gaussien et sinusoïdal, voir (Cressie, 1991)), quant à elle, ne permet pas de restituer la structure de Γ. La matrice Γ semble contenir beaucoup d'informations et sa paramétrisation par des structures simples dégrade la qualité du modèle contrairement à la paramétrisation des quantités α_1, α_0 et α_{-1}.

La validation en simulation de ce modèle révèle notamment via les fonctions de covariance spatio-temporelles que celui-ci capture la majeure partie de la structure spatio-temporelle des données : non-séparabilité et anisotropie nord-sud/ouest-est (voir Figure 1). Le modèle reproduit ainsi les déplacements moyens des évènements de conditions de vent liés au déplacements des masses d'air. La distribution marginale du processus est très bien reproduite par le modèle. Cependant, ce modèle ne permet de reproduire que les comportements

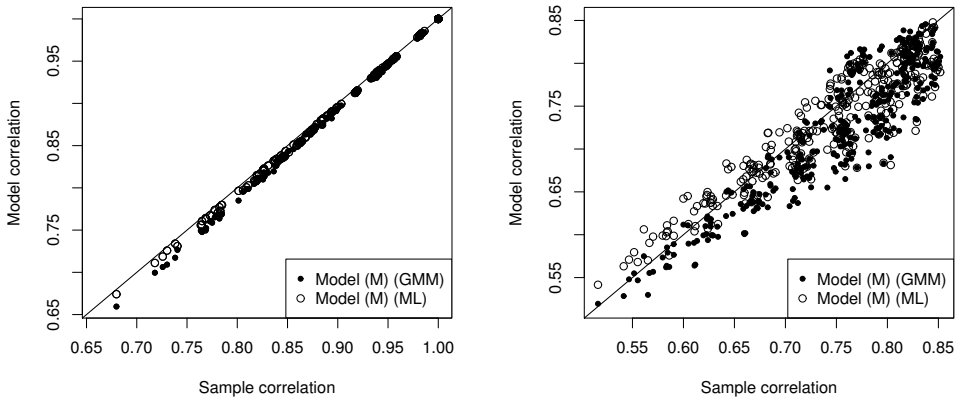

Figure 1: Corrélations estimées sur des simulations pour le modèle (M) en fonction des corrélations observées au temps 0 (gauche) et au temps 1 (droite). Les résultats sont présentés pour les deux méthodes d'estimation : GMM (méthode des moments généralisée) et ML (maximum de vraisemblance).

moyens des données sur la zone étudiée. Les données présentent les changements de régimes naturels liés au type de temps à grande échelle. Afin d'affiner la modélisation et de prendre en compte ces changements de régimes observés sur les données, nous considérons par la suite des modèles à changement de régimes.

Un modèle pour les composantes cartésiennes et polaires du vent en un site

Dans un second temps, nous nous intéressons aux processus des composantes polaires (U, Φ) et cartésiennes du vent (u, v) en un site afin de prendre en compte la direction du vent. Vitesse et direction du vent sont étroitement liées et la direction du vent est un bon descripteur des conditions météorologiques synoptiques. A notre connaissance, aucun modèle n'a été proposé pour modéliser jointement des séries temporelles de vitesse et direction du vent. Ceci peut être réalisé via la modélisation de (U, Φ) ou de (u, v). Dans (Holzmann et al., 2006), un modèle à variable latente est proposé pour modéliser un couple linéaire-circulaire de données. Cependant, conditionnellement à l'état caché, les variables linéaire et circulaire sont supposées indépendantes. Dans (Hering and Genton, 2010) un modèle de prédiction pour u et v est proposé, il est basé sur une régression linéaire avec une innovation de distribution Skew-t. Dans (Ailliot et al., 2006b), le couple (u, v) est modélisé par un modèle vecteur auto-

régressif à coefficients variant en fonction du déplacement du champ (u, v) entre deux instants successifs (voir aussi (Wikle et al., 2001; Modlin et al., 2012)).

Les données de vent présentent une alternance de périodes stables associées à une direction de vent généralement d'est et une intensité faible à modérée avec des périodes plus instables associées à des vents majoritairement d'ouest, plus volatiles et plus forts. Afin de restituer cette alternance et ainsi capturer l'information contenue dans les données au-delà du comportment moyen, nous proposons de considérer des modèles à changement de régimes. Ceux-ci ont été largement utilisés pour les données météorologiques : ils permettent de mieux restituer la dynamique temporelle des processus par rapport à des modèles sans régime (Ailliot and Monbet, 2012). Les modèles autorégressifs à changement de régimes cachés markoviens ont été initialement introduits pour les séries temporelles en économie dans (Hamilton, 1989). Ces modèles apparaîssent comme une généralisation des modèles à Chaîne de Markov Cachée et des modèles autorégressifs. Dans (Ailliot et al., 2006a) un modèle autorégressif à changement de régimes est proposé pour décrire l'évolution spatio-temporelle des champs (u, v).

Les transitions observées entre les régimes dépendent du vent observé, par exemple les transitions d'un régime dépressionnaire vers un régime anticyclonique se font en général lorsque la direction est associée à un vent venant du nord et sont très peu probables lorsque la direction est associée à un vent venant du sud. Dans le but de prendre en compte les observations passées, nous utilisons des probabilités de transition entre régimes non-homogènes similairement à (Hughes and Guttorp, 1994).

Nous considérons le modèle suivant : $(S_t)_{t\geq 0}$ est une chaîne de Markov à valeurs dans $\{1, 2, .., M\}$, S_t décrit le régime dans lequel se trouve l'observation Y_t à l'instant t. Cette variable est en général non-observée et nous la supposons ici latente. Notons pour un processus $\{X_t\}$: $X_t^{t+u} = (X_t, ..., X_{t+u})$ et $x_t^{t+u} = (x_t, ..., x_{t+u})$. Supposons que l'observation Y soit à valeurs dans $\mathbb{E}$.

Définition 1 *Soit $p, M \geq 1$ deux entiers, le processus $(S_t, Y_{t-p+1}^t)_{t\in\mathbb{Z}}$ est processus autorégressif à changement de régimes Markoviens cachés si c'est un processus de Markov à valeurs dans $\{1, ..., M\} \times \mathbb{E}$ et tel que :*

- *la distribution de S_t sachant $\{S_{t'}\}_{t'<t}$ et $\{Y_{t'}\}_{t'<t}$ ne dépend que de S_{t-1} et Y_{t-1}, on note $p_1(s_t|s_{t-1}, y_{t-1}) = P(S_t = s_t|S_{t-1} = s_{t-1}, Y_{t-1} = y_{t-1})$,*
- *la distribution conditionnelle de Y_t sachant $\{Y_{t'}\}_{t'<t}$ et $\{S_{t'}\}_{t'\leq t}$ ne dépend que de S_t et $Y_{t-1}, \ldots, Y_{t-p}$ et a pour densité de probabilité $p_2\left(y_t|s_t, y_{t-p}^{t-1}\right)$.*

Nous choisissons la paramétrisation de von Mises suivante pour p_1 avec la direction du vent Φ comme covariable :

$$p_1(s_t|s_{t-1}, y_{t-1}) \propto \Gamma_{s_{t-1},s_t} \exp(\kappa^{(s_{t-1},s_t)} \cos(\phi_{t-1} - \phi_0^{(s_{t-1},s_t)})), \tag{1}$$

où $\phi_0 \in [0, 2\pi]$, $\kappa \geq 0$ et Γ une matrice stochastique. Pour réduire le nombre de paramètres, le restriction suivante est considérée : $\kappa^{(s,s')} = \kappa^{(s')}$. De plus les contraintes suivantes d'identifiabilité sont appliquées : $\sum_{s'=1}^{M} \kappa^{(s')} = 0$.

Un modèle pour (u, v)

Nous introduisons un modèle MS-AR à émissions gaussiennes et à transitions non-homogènes décrites par p_1 pour le processus (u, v). Conditionnellement au régime S_t, l'observation Y_t s'écrit :

$$Y_t = B^{(S_t)} + A_1^{(S_t)} Y_{t-1} + A_2^{(S_t)} Y_{t-2} + ... + A_p^{(S_t)} Y_{t-p} + (\Sigma^{(S_t)})^{-1/2} \epsilon_t,$$

Y représente ici le processus bivarié (u, v), $B^{(i)}$ est un vecteur K-dimensionnel où K est le nombre de composantes de Y (ici $K = 2$), $A_1^{(i)}, ..., A_p^{(i)}, \Sigma^{(i)}$ sont des matrices $K \times K$ et ϵ est un bruit blanc gaussien de dimension K. Les transitions de la chaîne S sont définies par p_1 en (1). L'étude des critères BIC et de statistiques calculées sur les données et sur des échantillons simulés conclut à choisir un modèle avec $M = 3$ et $p = 2$. Le graphe des distributions conditionnelles est le suivant, pour $p = 1$:

$$\begin{tikzcd}
\cdots \arrow[r] \arrow[d] & S_{t-1} \arrow[r] \arrow[d] & S_t \arrow[r] \arrow[d] & S_{t+1} \arrow[r] \arrow[d] & \cdots \arrow[d] \\
\cdots \arrow[r] \arrow[ur] & Y_{t-1} \arrow[r] \arrow[ur] & Y_t \arrow[r] \arrow[ur] & Y_{t+1} \arrow[r] \arrow[ur] & \cdots
\end{tikzcd}$$

Un modèle pour (U, Φ)

Premièrement, nous ajustons un modèle MS-AR non-homogène à la variable Φ, p_2 est défini par un processus de von Mises (Breckling, 1989) et p_1 est défini en (1). Pour ce modèle nous choisissons $M = 4$ et $p = 2$. L'intensité du vent U est ensuite modelisée par un modèle MS-AR non-homogène avec p_1 paramétré comme ci-dessus. Sous les mêmes critères que précédemment le modèle est choisi avec $M = 3$ et $p = 2$. Le graphe des distributions conditionnelles dans

ce modèle pour (U, Φ) est le suivant, pour $p = 1$:

$$\begin{tikzcd}
\cdots \arrow[r] \arrow[d] & S_{t-1}^{(\Phi)} \arrow[r] \arrow[d] & S_t^{(\Phi)} \arrow[r] \arrow[d] & S_{t+1}^{(\Phi)} \arrow[r] \arrow[d] & \cdots \arrow[d] \\
\cdots \arrow[r] \arrow[ur] \arrow[dr] & \Phi_{t-1} \arrow[r] \arrow[ur] \arrow[dr] & \Phi_t \arrow[r] \arrow[ur] \arrow[dr] & \Phi_{t+1} \arrow[r] \arrow[ur] \arrow[dr] & \cdots \\
\cdots \arrow[r] \arrow[d] & S_{t-1}^{(U)} \arrow[r] \arrow[d] & S_t^{(U)} \arrow[r] \arrow[d] & S_{t+1}^{(U)} \arrow[r] \arrow[d] & \cdots \arrow[d] \\
\cdots \arrow[r] & U_{t-1} \arrow[r] & U_t \arrow[r] & U_{t+1} \arrow[r] & \cdots
\end{tikzcd}$$

Résultats

Dans chaque cas, le modèle à transitions non-homogènes restitue plus précisement les quantités étudiées que le modéle avec des transitions homogènes entre les régimes. Ces modèles MS-AR non-homogènes révèlent de bonnes capacités à reproduire les distributions jointes et marginales de (u, v) et (U, Φ) et la structure d'ordre 2 de ces processus. De plus, la comparaison entre les conditions simulées de (u, v) et de (U, Φ) révèle que la dynamique temporelle du vent est mieux reproduite par le modèle pour les coordonnées cartésiennes que par le modèle pour les coordonnées polaires. Les distributions jointe et marginales de (u, v) tendent à être reproduites de manière équivalente par les deux modèles. D'autres statistiques sont également considérées comme le nombre de rotations dans le sens horaire et anti-horaire, le modèle pour (U, Φ) semble le mieux décrire cette statistique (voir Figure 2).

Modèles multi-sites pour les composantes cartésiennes du vent

Nous étudions dans la troisième partie une extension du modèle homogène MS-AR pour (u, v) au cadre multi-site. Cette extension soulève la question de la pertinence d'un régime régional commun à tous les sites. Nous montrons que cette hypothèse est raisonnable si la zone choisie est suffisamment homogène. En effet l'ajustement d'un modèle MS-AR homogène en plusieurs sites montre une cohérence entre les régimes déterminés en chaque site et entre les coefficients des modèles autorégressifs dans chaque régime.

Nous proposons également plusieurs modèles autorégressifs multi-variés à changement de régimes observés. Un de nos objectifs est de comparer ces modèles selon que les régimes sont observés ou latents. Ces deux types

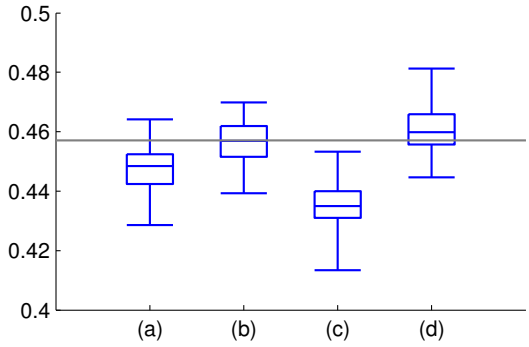

Figure 2: Frequence de rotations anti-horaires entre deux observations pour les différents modèles. La ligne grise correspond à la valeur obtenue sur les observations (45.4 % de rotations anti-horaires contre 54.6 % de rotations horaires). Modèles : (a) MS-AR homogène pour (u, v), (b) modèle homogène pour (U, Φ), (c) MS-AR non-homogène (u, v), (d) modèle non-homogène pour (U, Φ).

de modèles ont largement été utilisés pour modéliser les variables météorologiques. Nous proposons des régimes *a priori* déterminés à partir d'une variable atmosphérique grande échelle : la hauteur de géopotentiel à 500mb pour l'Atlantique Nord-Est et déterminés à partir des variables locales étudiées à l'aide de méthodes de classification adaptées. Nous discutons ici les méthodes de classification et le choix des descripteurs afin de construire une classification observée pertinente du point de vue météorologique et du point de la validation en simulation. Nous montrons ici que la classification extraite sur la hauteur de géopotentiel ne permet pas d'extraire des régimes aussi marqués que lorsque la classification faite sur les données locales. La classification *a priori* la plus adaptée semble celle extraite par un modèle à chaîne de Markov cachée à émission gaussiennes sur les données $\{u_t - u_{t-1}, v_t - v_{t-1}\}$. Les régimes provenant de modèles à changement de régimes cachés présentent quant à eux une certaine optimalité vis-à-vis du modèle et des données. Cependant ceux-ci ne contiennent pas d'information sur la circulation atmosphérique à une échelle plus grande que celle de la zone étudiée.

De plus nous décrivons le lien entre les régimes déterminés à petite échelle spatio-temporelle, avec les régimes de temps à grande échelle observés en Atlantique Nord. L'alternance de ces derniers gouverne la météorologie de l'Europe, il est donc d'intérêt d'étudier leur influence sur les régimes à plus petite échelle déterminés à partir des données de vent. Nous montrons que les régimes petite échelle apparaîssent dans des régimes grande échelle privilégiés.

Lorsque les régimes *a priori* sont choisis pertinemment, les deux types de modèles tendent à se comporter de manière équivalente en simulation. Notamment nous mettons en évidence le bénéfice des modèles à changement de

régimes dans la description de l'alternance en intensité et variabilité des conditions de vent.

Conclusion

Nous avons proposé ici plusieurs modèles pour simuler des données de vent à l'echelle régionale au large de la Bretagne. Un premier modèle permet de restituer les déplacements moyens des conditions de vent liées aux déplacements des masses d'air. L'originalité de ce modèle est l'introduction d'une variable latente représentant le vent à l'échelle régionale. Le vent local est ensuite obtenu par projection linéaire de cette condition régionale. Deux perspectives d'amélioration de ce modèle sont possibles afin d'affiner la modélisation temporelle. Une possibilité est d'introduire des changements de régimes afin de proposer une description des données à une échelle plus fine que l'échelle moyenne. Cependant l'inférence d'un modèle à deux variables latentes est très complexe. Une autre perspective est de modéliser le vent moyen régional par un processus multi-dimensionnel afin de capturer plus d'information contenue dans les données.

Dans la suite de ce travail, nous avons construit deux modèles pour les séries temporelles de conditions de vent prenant en compte la dépendance jointe entre direction et intensité du vent. En l'état actuel de nos connaissances, aucun générateur de ce type n'avait été proposé avant ce travail. Ces deux modèles permettent de simuler les coordonnées polaires et cartésiennes du vent en un site. Les deux modèles proposés font intervenir une structure de changement de régimes cachés dont les transitions sont gouvernées par la direction du vent. Nous avons ici comparé ces deux modèles en terme de réalisme des séquences simulées. Notamment, le modèle pour les composantes cartésiennes du vent tend à mieux reproduire la dynamique temporelle des conditions de vent tandis que les distributions jointe et marginales de (u, v) tendent à être reproduites de manière équivalente par les deux modèles.

Pour finir, nous présentons un générateur multi-site de coordonnées cartésiennes du vent. L'originalité de ce modèle est qu'à notre connaissance, aucun générateur multisite n'avait été conçu pour les coordonnées cartésiennes du vent. De plus la question de la modélisation du régime dans un cadre multisite n'a pas été traitée. Une perspective serait de construire une procédure de test basée sur le rapport de vraisemblance permettant de décider la pertinence d'un régime régional ou local. Nous avons également proposé une comparaison en simulation et en terme d'interprétation météorologique des modèles à changement de régimes cachés et observés. Ces deux types de modèles ont été très largement utilisés sans avoir été comparés. Lors de cette comparaison spécifique aux données de vent, nous avons mis en évidence la difficulté d'extraire une classification à la fois pertinente météorologiquement et en terme de de-

scription du modèle conditionnel pour ces données. Nous avons également mis en évidence que l'apparition des régimes à petite échelle se fait dans des régimes à grande échelle privilégiés. Par la suite, une paramétrisation adaptée des matrices autoregressives permetterait d'ajuster le modèle sur de plus grands jeux de données en évitant les phénomènes de sur-paramétrisation.

Dans cette thèse, nous avons considéré plusieurs modèles à variable latente. Un premier modèle fait intervenir une variable latente à valeurs continues. L'identifiabilité de celui-ci a été étudiée et deux méthodes d'estimation de ce modèle ont été comparées. Un second type de modèle étudié est celui des modèles à variable latente discrète. Nous avons ici proposé une modélisation originale des transitions entre les valeurs de cette variable. En terme d'applications, des modèles multi-sites pour les données de vent ont été proposés, un premier permet de restituer les déplacements moyens des masses d'air. La modélisation de la loi jointe et marginales du processus des coordonnées cartésiennes du vent a été étudiée. Enfin une prise en compte des régimes observés sur les données a été réalisé via les modèles à changement de régimes. Nous avons également montré que ceux-ci sont influencés par les régimes de temps à grande échelle qui régissent la météorologie en Europe.

Chapter 1

Introduction and context

Those last decades, more and more impact studies involve meteorological measures and simulations, for instance in the growing field of wind energy. Various measures related with wind are needed, such as short-term wind power production or weather uncertainties for the maintenance of off-shore wind farms. Furthermore, according to the kind of applications, one may need a large number of long sequences of realistic wind data, such as in impact studies. Stochastic generators of artificial sequences of weather conditions have been introduced to respond this problem. They are statistical models that are calibrated on a dataset and that aim at simulating sequences of meteorological variables with statistical properties similar to the ones of the calibration set. They can also be used as statistical downscaling tools.

Meteorological time series exhibit non-linearities induced by various causes such as space-time non-separability, regime-switching patterns or interaction with other variables. Meteorological data may be discrete-valued, continuous or circular and interaction may occurred between these variables of various natures. Describing these features is very challenging from a statistical point of view. Indeed the modeling has to be sophisticated enough to handle these patterns but tractable to keep the estimation and simulation feasible. We propose in this work various approaches to build space-time models that handle the non-linearities of the studied wind direction and wind speed time series.

In this work, we propose to construct stochastic generators of surface wind conditions off-shore Brittany in France. We base the construction of wind conditions generators on Hidden Markov models and state-space models, which both include a latent process that represents a non-observed regional-scale component. In a first time a Gaussian linear state-space model shows good abilities to reproduce average behaviors of wind speed. To go one step ahead and model the scale beyond the average one and to consider all the information of wind fields, we use regime-switching models for polar and Cartesian coordinates of wind. In this modeling, a discrete variable describes the current weather type and helps to model the regime-switching observed on data due

to the large-scale meteorological conditions.

The introduction is organized as follow, we present a brief state of the art about stochastic weather generators in Section 1.1. In Section 1.2, we describe sources of meteorological data, the dataset under study and its associated modeling challenges. In Section 1.3, we describe the classes of Hidden Markov models and state-space models that we use to build stochastic generators that encompass the features of the data. The following sections are linked to the patterns observed on the data, which we want to reproduce. Namely in Section 1.4, we describe the way of accounting space-time motions for multi-site dataset of time series and accounting for regime-switching. In Section 1.5, we describe the literature on existing models dealing with interactions between wind direction and speed. We finish in Section 1.6 by a description of the proposed work.

1.1 Stochastic weather generators

A growing number of recent impact studies requires a large number of long sequences of meteorological data at fine spatial scale that are consistent with the observed meteorology and climatology, such as in hydrological design, in agricultural or ecosystem simulation. In that purpose, stochastic weather generators have been developed, namely they simulate, in a statistical framework, unlimited number of sequences of meteorological data at small spatial and temporal scales. Those sequences may account for climate change scenarios and enable to reproduce a greater part of the variability of the considered process that may not be assessed through a limited number of observed sequences.

The major fields of applications of the stochastic weather generators are the followings. These models have been adopted in various impact studies as a computationally inexpensive tool that generates quickly as many synthetic time series as desired of unlimited length without missing data (Flecher et al., 2010). Stochastic weather generators can also be used as a tool to conditional simulation of missing values (Yang et al., 2005), besides parameters of stochastic weather generators can be spatially interpolated to generate meteorological data at non-observed station (Wilks, 1998; Kleiber et al., 2012). Concerning the third field of application, these models have been combined to downscaling techniques to generate scenarios of climate change at a local scale (Semenov and Barrow, 1997) or they can be used to generate synthetic data at small spatio-temporal scale consistent with climate changes scenarios at larger spatio-temporal scale (Wilks, 1992), see (Maraun et al., 2010; Wilks, 2010, 2012) for the use of stochastic weather generators in downscaling.

Two approaches are used to build stochastic weather generators: the empirical one based on re-sampling, non-parametric methods (Rajagopalan and Lall, 1999) and model-based methods (Richardson, 1981; Wilks, 1999; Flecher

et al., 2010). We focus on this second method in this thesis. It has the advantage of bringing a framework that leads to easy interpretations and enables to simulate unobserved events. However more efforts have to be paid on the modeling than when considering non-parametric models. When constructing a stochastic weather generator, one has to find a good compromise between the simplicity of the model, that ensures an ease of fitting and simulation, and the fidelity to the data.

In the early days of stochastic weather generators, most of the efforts were focused on precipitation processes at one location, see (Richardson, 1981) or (Wilks and Wilby, 1999; Maraun et al., 2010; Srikanthan and McMahon, 1999) for a review. This meteorological variable is of great interest for applications and many generators simulate other variables conditionally to the rain state. Modeling several meteorological variables is challenging since the dependence between variables of various nature has to be accounted. Various single-site models of multivariate meteorological dataset have been proposed (Wilks, 1999; Flecher et al., 2010). Recent researches are focused on the development of multi-site models, which is a challenging issue especially in the context of multivariate meteorological variables (Wilks, 2009; Kleiber et al., 2013).

Wind generators have in particular been used to assess wind power production (Brown et al., 1984), to account for coastal erosion (Skidmore and Tatarko, 1990), drift of objects in the ocean (Ailliot et al., 2006a) or weather uncertainties into a simulator of maintenance costs of off-shore wind-farm (Hofmann and Sperstad, 2013). Various stochastic weather generators of multiple variables include wind speed modeling, however specific wind generators have been developed to refine the modeling (Ailliot and Monbet, 2012). To the best of our knowledge, very few multi-site models for wind conditions have been proposed and especially for Cartesian coordinates of wind.

1.2 Meteorological data

1.2.1 Various sources of data

In situ wind data may come from direct measures from inland stations, from off-shore buoys or from satellites. Data may contain missing values or exhibit non-regular sampling (especially satellite data). Multi-site sets of time series are often non-consistent in space and time or are available over periods of time that are not long enough to estimate reliably probabilities of complex events. Besides one may have reservation about the quality of observational data due to the possible bias or aging of sensors, changes of measurement site. Homogenization of observed meteorological data is a current research topic, however there exists no standard methods to homogenize wind data due to their important variability in space and time. Ground measurements of wind

is subject to topography that has a strong influence on this process

As a surrogate to observed data, one may use reanalysis data that fill gaps and provide estimates of unobserved data. They are obtained by combining observations and Numerical Weather Prediction models; and provide a historical data base, which is available at a regular temporal and spatial resolution. However they provide a limited number of sequences of meteorological processes that are coherent with the experienced meteorology and climatology.

Many recent investigations tend to use data consistent with possible future climate. Preferred source of climate projections for impact studies is Global Climate Models (GCM), these models based on physics perform reasonably in simulating the current climate over regional and global scale. However these models are computationally expensive and are not reliable at fine spatial scale due to their course spatial resolution. Downscaling tools have been developed to fill the gap between the resolution required for end users and the climate change scenarios. One approach is the dynamical downscaling where outputs of a GCM are used to drive a Regional Climate Model (RCM) at a higher spatial resolution. Another approach is the statistical downscaling where statistical links are established between local observations and large-scale variables.

1.2.2 Considered wind data and associated issues

In this subsection, we introduce the data we use and the associated modeling challenges in the context of stochastic weather generators. In situ data are neither available on a long time period nor on a large area offshore Brittany in France. Besides the quality of various dataset is poor due to the lack of homogenization. In this work, we consider wind data (zonal and meridional components (u, v)) at 10 meters above sea level extracted from the ERA Interim Full dataset produced by the European Center of Medium-range Weather Forecast (ECMWF). The dataset can be freely downloaded and used for scientific purposes at the URL *http://data.ecmwf.int/data/*. This dataset is available on a regular space-time grid with a temporal resolution of 6 hours and a spatial resolution of $0.75°$. The observed wind fields are generally smooth, which leads to a high correlation between the different sites. Although the smoothness observed here is inherent to reanalysis data that are known to be smoother than observations, it is also coherent with the considered spatio-temporal scale. A convention when using wind direction is that we deal with the direction from where the wind is blowing rather than with the direction to which it blows.

Off-shore the coasts, without topographical obstacle, reanalysis data are a good approximation of observational data. In a first step, in order to settle the modeling in a simple context, we consider a rectangular area away from the coastline to avoid local effects due to the coasts. Our purpose is then to consider data along the coast which may be more useful for the applications. However the methodology introduced in this paper could easily be adapted to

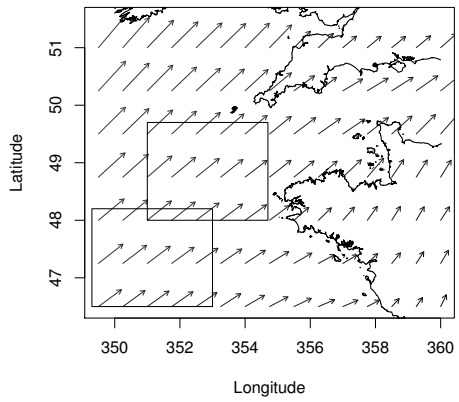

Figure 1.1: Map of the two studied areas. Arrows correspond to the average field of (u, v).

handle datasets with more complicated space-time sampling such as the one obtained when considering networks of meteorological stations.

The dataset consists in 33 years of wind data from 1979 to 2011 and we focus on wintertime to avoid seasonal effects. Daily components are weak in wintertime and are neglected in the following. Further, the statistical inference is based on the assumption that the wintertime months of wind data are independent realizations of a common stationary stochastic process. This assumption is usual for meteorological processes but it does not take into account low frequency variations such as the North Atlantic Oscillation (NAO). In a first time, the most northern area was studied, later we found that the bottom rectangle is more homogeneous in terms of wind events and then we worked with this area, see Figure 1.1.

Challenges raised while considering this dataset and more generally multi-site wind data are developed in the following paragraphs. They are linked with the distributions and dependence of the various components of wind fields and with the description of space-time propagation of wind events and the observed alternations between stable wind conditions and more volatile wind conditions. In the sequel wind intensity is denoted as U, wind direction as Φ and zonal and meridional components of wind as u and v.

• *Marginal distribution and complex dependences at one site.*
· *Joint and marginal distributions.* In Figure 1.2, distributions of wind speed U and of the components (u, v) are depicted. Wind speed distribution is skewed and the joint distribution of (u, v) admits complex patterns. The margin of u reveals two separate modes whereas the one of v does not exhibit

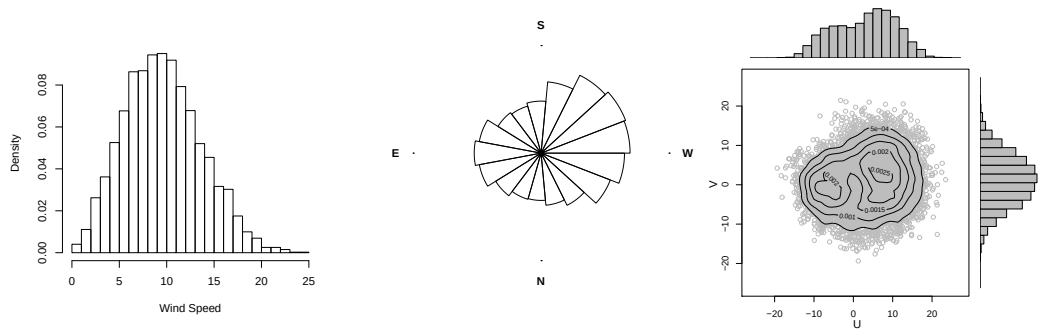

Figure 1.2: Left: histogram of wind intensity data, central panel: rose diagram of wind direction, right: joint and marginal distribution of (u, v) at the location $(47.25°N, 9.75°W)$.

a clear bi-modality. The bi-modality of marginal distributions urges to use models with regime-switching. The very few points around the point $(0, 0)$ indicate 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. Power transformations are commonly used to approximate Gaussian margins. In order to handle the asymmetry of the wind speed distribution, the Box-Cox transformation is applied, see for instance (Brown et al., 1984). More precisely, let us denote for a given $\lambda_i \geq 0$,

$$\begin{cases} U_{\lambda_i,i,t} &= \frac{U_{i,t}^{\lambda_i}-1}{\lambda_i} \quad \text{if } \lambda_i > 0 \\ U_{\lambda_i,i,t} &= \log(U_{i,t}) \quad \text{if } \lambda_i = 0. \end{cases}$$

with $U_{i,t}$ the wind speed at time t and location i. Following (Hinkley, 1977), λ_i can be estimated by searching the roots of the asymmetry measure

$$S(\lambda_i) = \frac{\text{mean}(U_{\lambda_i,i,t}) - \text{median}(U_{\lambda_i,i,t})}{\sqrt{\text{var}(U_{\lambda_i,i,t})}}. \tag{1.1}$$

In Chapter 2, the same power transformation is applied at each site to preserve the variance structure. The average value of the $\hat{\lambda}_i$ estimated at each site is used as a common power: $\hat{\lambda} = 0.85$. In Chapter 2, the model is fitted on these transformed data.

The following transformation is used on both components u and v:

$$\begin{cases} \tilde{u} &= U^{\alpha}\cos(\Phi) \\ \tilde{v} &= U^{\alpha}\sin(\Phi). \end{cases}$$

This transformation with $\alpha > 1$ aims at filling the hole around $(0, 0)$ in order to facilitate the modeling. In practice, α is chosen empirically equal to 1.5. In

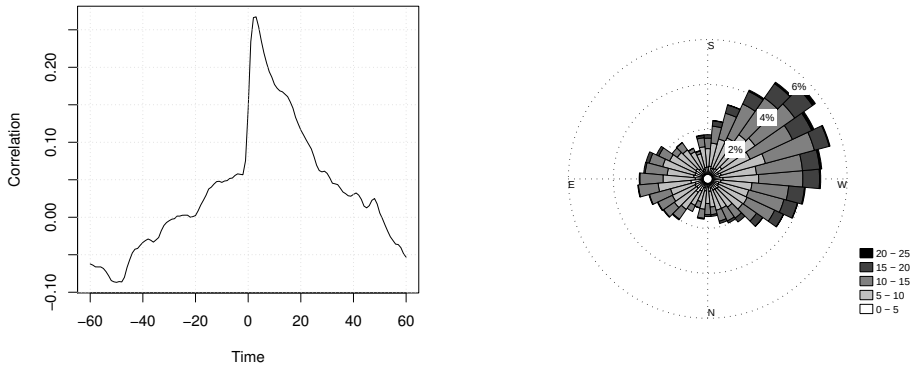

Figure 1.3: Left: cross-correlation between u and v. Right: wind rose plot of wind speed and direction, at the location $(47.25°N, 9.75°W)$.

Chapters 3 and 4, we will see that the description of this distribution by the proposed models is satisfying.

The value of $\hat{\alpha}$, which is greater than 1, enables to fill the hole around zero by strengthening the values of weak winds. Whereas the value of $\hat{\lambda}$, which is smaller than 1, allows to reduce the positive skewness of wind speed distribution by reducing the transformed values of strong winds.

The circular nature of wind direction requires to use specific distribution probability models. Most of the proposed models are based on von Mises and wrapped normal distributions that have a as central role for circular data as the normal distribution has for linear variables. In Chapter 3, we propose to work with von Mises distribution to handle the circular nature of wind direction, see (Breckling, 1989).

· *Intensity and direction interaction.* A typical pattern of wind fields is the dependence between speed and direction, which is also observable on (u, v) data. Prevailing flows are westerly in wintertime, this induces a temporal advance of the meridional component v on the zonal one u, see Figure 1.3. In the right panel of Figure 1.3, one can notice the strong link between intensity and direction of wind, stronger wind conditions are more likely to occur in westerly blowing conditions and in easterly conditions, intensity is weaker. This phenomenon is also observable on the time series of Figure 1.4. To cope with this dependence, the MS-AR models that we introduce in Chapters 3 and 4 are designed for polar and Cartesian coordinates of wind. Moreover, transitions between stable and volatile periods generally occur in privileged wind direction, for instance transitions from unstable conditions to stable ones are more likely when wind is southward. We propose in Chapter 3 to account

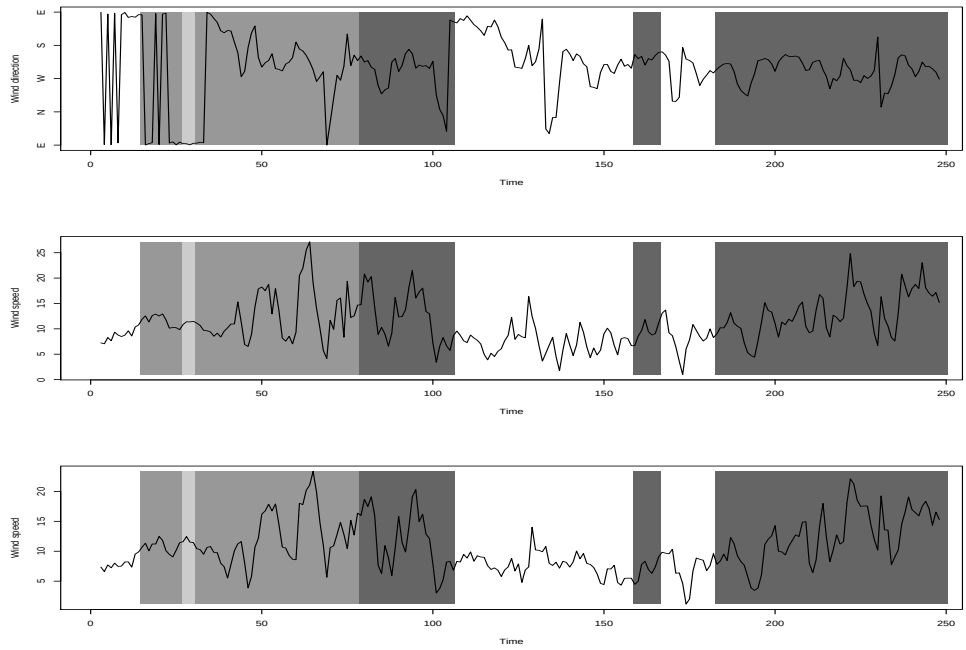

Figure 1.4: Top and central panels: wind direction and speed at a western location $(48°N, 10.5°W)$, bottom panel: wind speed at an eastern location $(48°N, 6.75°W)$. Weather regimes in grey, from the lighter to the darker: BL, AR, NAO+, NAO-.

for that pattern by driving the transitions between regimes by wind direction.

• *Space-time motions and non-separability.* In wintertime, the prevailing air masses are generally moving westerly. In Figure 1.4, we can observe the propagation of maximal wind speed from a western site (central panel) to an eastern one (bottom panel). It induces non-separability between time and space components of the space-time covariance function of wind processes. The asymmetry with respect to 0 of lagged by 1 cross-correlations of wind speed shown in Figure 2.2 highlights this phenomenon. Moreover, one can notice on this figure some anisotropy patterns. Indeed dependences in latitude and longitude differ. Methods discussed in Subsection 1.4.2 generally enable to reproduce these average patterns. We introduce in the modeling two scales to capture these patterns, a regional scale, which is not observed, is modeled with its own dynamic as a latent process and conditionally to it a model drives the local scale. In Chapters 2 and 4, we show that the proposed model enables to capture and reproduce space-time interactions of wind and to deal with the spatial non-stationarity.

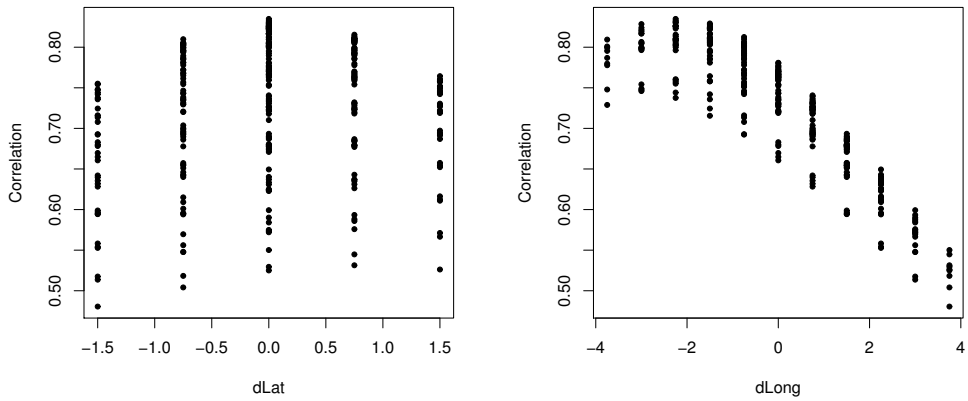

Figure 1.5: Lagged-one cross-correlations of wind speed against differences of latitude (left) and longitude (right).

• *Regime-switching patterns.* The meteorology and climatology of Europe are governed in part by the alternate of large-scale weather regimes (Michelangeli et al., 1995; Cassou, 2008). These are characteristic patterns of the atmospheric circulation above the North-Atlantic Ocean. In winter, four regimes are identified and described in various references (Michelangeli et al., 1995; Cassou, 2008; Najac, 2008). They correspond to the two phases of North-Atlantic Oscillation (NAO+ and NAO-), the blocking (BL) and the Atlantic Ridge (AR) regimes. In France in wintertime, these four regimes respectively correspond to privileged flows that are respectively: south-western flows with numerous storms (NAO+), western slow flows (NAO-), southern or eastern very stable flows (BL) and northern flows (AR).

The large-scale conditions influence the local wind, it is then observed an alternation of different intensity and variability of wind conditions. In Figure 1.4, we can see that volatile conditions are associated with the NAO+ phase, whereas more stable wind conditions are associated with BL and AR regimes. In Figure 1.4, one can also notice that wind has stronger intensity and temporal variability when wind is westerly, whereas easterlies are generally associated with lower and more stable wind conditions. Regime-switching models are introduced in Subsection 1.4.3 to reproduce this instantaneous alternate of different temporal variabilities and intensities in wind conditions. In Chapters 3 and 4, we propose Markov-Switching Autoregressive (MS-AR) models in which the current weather state is not observed and then modeled as a hidden Markov chain. In Chapter 4, we also discuss the choice of a latent or observed weather state.

1.3 Hidden Markov and state-space models

In this section, we describe the general classes of models that we use to cope with the various features of the data in our aim to construct a realistic stochastic wind conditions generators. Namely in this work, we propose a Gaussian linear state-space models for wind speed and hidden Markov-Switching Autoregressive models for polar and Cartesian components of wind. Hidden Markov models and state-space models have been extensively used in many domains. They bring a very flexible framework for modeling time series (see (Zucchini and MacDonald, 2009; Durbin and Koopman, 2012; Brockwell and Davis, 2002)) and space-time processes (Wikle and Hooten, 2010).

1.3.1 Basic properties

The idea of hidden Markov and state-space models is that the behavior of the system is determined by an unobserved series $\{X_t\}_t$ with which are associated a series of observations $\{Y_t\}_t$. The relation between the unobserved time series and the observed one is specified by conditional independence assumptions. The most basic conditional structure of these models is represented by the following Directed Acyclic Graph (see (Durand, 2003) for additional information about DAGs):

$$\begin{tikzcd}
\cdots \arrow[r] \arrow[d] & X_{t-1} \arrow[r] \arrow[d] & X_t \arrow[r] \arrow[d] & X_{t+1} \arrow[r] \arrow[d] & \cdots \arrow[d] \\
\cdots & Y_{t-1} & Y_t & Y_{t+1} & \cdots
\end{tikzcd}$$

When X takes discrete values, these models are denoted Hidden Markov Models (HMM) whereas when X is continuous-valued, the term of state-space models is used.

• *Gaussian linear state-space model.* Gaussian linear state-space models have been widely used in engineering and control theory since they provide a simple and flexible framework. They are written as:

$$\begin{cases} X_{t+1} = \rho X_t + \Sigma^{1/2}\epsilon_{t+1} & (1.2) \\ \quad Y_t = \Lambda X_t + \Gamma^{1/2}\eta_t \quad \text{for } t \geq 0, & (1.3) \end{cases}$$

$X_t \in \mathbb{R}^d$ and $Y_t \in \mathbb{R}^K$, $\{\epsilon_t\}$ and $\{\eta_t\}$ are independent Gaussian white noises with zero-means and identity covariance matrices. The autoregressive parameter ρ is a $d \times d$-matrix. The loading $K \times d$-matrix Λ links the hidden state and the observation, Σ and Γ are covariance matrices of dimension $d \times d$ and $K \times K$, which model respectively the structure of the innovation of the latent state and the observation error. The equation (1.2) is called the state

equation and (1.3) refers to the observation equation. This framework can be generalized to a non-linear and non-Gaussian context, which necessitates more complex inference procedures.

The strong correlation of wind conditions between sites urges the use of a common signal to all the locations, which is interpreted as the regional wind in the sequel. In Chapter 2, the regional wind is explicitly introduced as a latent variable X, with its own autoregressive dynamic, and the local wind Y is expressed as a function of the regional wind at different lags to model the mean displacement of the air masses.

• *Hidden Markov-Switching AutoRegressive Models.* Markov-Switching AutoRegressive (MS-AR) models appear as a generalization of Hidden Markov Models (HMM) in allowing temporal dynamics within the regimes (Hamilton, 1989). Let denote $Y_t^{t+u} = (Y_t, ..., Y_{t+u})$, for $t > 0$ and $u > 0$, this notation holds true with other variables. Let $p, M \geq 1$ be some integers, the sequence $(X_t, Y_{t-p+1}^t)_{t\in\mathbb{Z}}$ follows a MS-AR model if it is a Markov chain with values in $\{1, ..., M\} \times \mathbb{E}$ 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 on 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 $Y_{t-1}, \ldots, Y_{t-p}$.

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

$$\begin{tikzcd}
\cdots \arrow[r] \arrow[d] & X_{t-1} \arrow[r] \arrow[d] & X_t \arrow[r] \arrow[d] & X_{t+1} \arrow[r] \arrow[d] & \cdots \arrow[d] \\
\cdots \arrow[r] & Y_{t-1} \arrow[r] & Y_t \arrow[r] & Y_{t+1} \arrow[r] & \cdots
\end{tikzcd}$$

One generalization of hidden MS-AR models is the case of models where transitions are driven by the past observation of Y. In this case, 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 on Y_{t-1}. In the sequel these models are referred to non-homogeneous MS-AR models and are used in Chapter 3 to improve the modeling of regimes. The associated directed graph is written as, for $p = 1$:

$$\begin{tikzcd}
\cdots \arrow[r] \arrow[d] & X_{t-1} \arrow[r] \arrow[d] & X_t \arrow[r] \arrow[d] & X_{t+1} \arrow[r] \arrow[d] & \cdots \arrow[d] \\
\cdots \arrow[r] \arrow[ur] & Y_{t-1} \arrow[r] \arrow[ur] & Y_t \arrow[r] \arrow[ur] & Y_{t+1} \arrow[r] \arrow[ur] & \cdots
\end{tikzcd}$$

In Chapters 3 and 4, in order to capture consistent instantaneous behaviors linked with the current weather state and especially the alternate of temporal

variability of wind conditions, we consider models with regime-switching. The weather type is described by a hidden Markov chain which represents the current weather state over the considered area. Conditionally to the weather state, the local scale is modeled using an autoregressive model.

1.3.2 Inference

One typical aim of the statistical inference is to infer about the hidden state X and about the parameters of the model. One of the main advantage of these two classes of models is that estimation, forecasting and smoothing can be processed through general and efficient procedures.

• *Identifiability.* The introduction of a latent process X is a source of non-identifiability of these models since parameters of the whole system have to be estimated only via the observations Y. In practice the non-identifiability is a source of numerical instability, its study is then of great importance. The identifiability of the proposed models has been studied in the literature under various points of view. Identifiability of linear Gaussian state-space models was initially investigated in control theory and has been largely explored during the last decades (Hannan and Deistler, 1988; Ljung, 1999; Bai and Wang, 2012; Bork, 2010). Concerning the identifiability of HMM models, see (Cappé et al., 2005) and references therein and for the identifiability of hidden MS-AR models see (Francq and Roussignol, 1998; Krishnamurthy and Ryden, 1998).

• *Prediction, filtering and smoothing.* When inferring the hidden state, one can consider three procedures: prediction, filtering and smoothing. The goal of filtering (respectively smoothing, respectively prediction) is to obtain as much as possible information about the hidden variable X_t from the observations $(y_1, ..., y_t)$ (respectively $(y_1, ..., y_T)$, respectively $(y_1, ..., y_{t-1})$). The solution consists in computing recursively the conditional distribution of X_t according to $(y_1, ..., y_t)$ (respectively $(y_1, ..., y_T)$, respectively $(y_1, ..., y_{t-1})$), which realizes the best approximation of X_t according to $(y_1, ..., y_t)$ in terms of a chosen error. When considering a Gaussian linear state-space model, this inference is performed through the Kalman recursions. Regarding hidden Markov-Switching AutoRegressive models, this is performed by the Forward-Backward recursions (Hamilton, 1989).

• *Expectation-Maximization algorithm.* The parameters estimation of models with latent variables is generally performed via the Expectation-Maximization (EM) algorithm which proceeds into cycling through the two following steps (Dempster et al., 1977). The EM-algorithm aims at maximizing the complete log-likelihood function based on the observations Y, assume that y_{-1} and y_0 are observed:

$$\theta \to \mathrm{E}(\log(\mathcal{L}(\theta; Y_1, ..., Y_T, X_1, ..., X_T))|Y_{-1}^T = y_{-1}^T),$$

it is proven that through the iterations of the algorithm, a convergent sequences of approximation of the Maximum Likelihood estimator of θ is computed. The Expectation-step enables to update the values of the expected complete likelihood that involves the conditional distribution of X according to Y under the current value of the estimate parameters. Smoothing techniques are used to that aim. The Maximization-step consists in updating the parameter estimates by maximizing the precedent value of the expected complete likelihood. Explicit expressions of the optimal parameters can be available like for Gaussian linear state-space model, or numerical procedures may be needed.

In this work, we introduce a Gaussian linear state-space model and various hidden MS-AR models to encompass the various features of the data described in Section 1.2. The following sections are organized according to the various topics linked with the features that we want to reproduce. In each following section, we give a small state of the art of techniques and models related to these topics and we propose our associated modeling.

1.4 Modeling space-time dependence

One of our interests is the ability of the proposed model to describe and reproduce space-time motions of wind events. Meanwhile, the marginal temporal dynamic of the data is also to be described accurately. Various models have been proposed for wind speed at one site however very few models have been constructed for multi-site wind time series. Nevertheless, a wide variety of models and techniques are suitable to account for space-time propagation such as state-space models. In the following, we present briefly and non-exhaustively methods to account for the temporal dependence and space-time patterns in parametric stochastic weather generators. Then we introduce regime-switching models, which we use to capture instantaneous behaviors of the data.

1.4.1 Temporal dependence at a single site

First weather generators were dealing with precipitation occurrences, a simple approach to account for the temporal dynamic of this kind of data is the framework of finite state space Markov Chain (Gabriel and Neumann, 1962). One step further is the modeling of continuous valued time series, which is realized through linear autoregressive (AR) models (Richardson, 1981). This classical approach for modeling continuous valued time series at a single location consists in using the Box-Jenkins methodology, where an AR model (or more generally an ARMA model) is fitted after achieving stationarity and applying a marginal transformation to obtain Gaussian like margins. This method has been widely used for wind time series and the most usual transformation of

wind data is a power transformation (Brown et al., 1984; Nfaoui et al., 1996; Kamal and Jafri, 1997), but specific distributions are used as well, for instance Weibull (Brown et al., 1984). Another approach similar to the auto-regressive modeling but with more flexibility, consists in specifying the conditional distribution of the variables between two consecutive times like in (Flecher et al., 2010), where a closed skew normal distribution is used. Most of these models enable to describe in a satisfying way the average and short-term behaviors of the data. In order to model more than the average behaviors, one can introduce non-linear models among them Artificial Neural Networks (ANN), models with latent variables as state-space models, described in Subsection 1.3, or models with regime-switching described in Subsection 1.4.3. They have proven their ability to improve the temporal dynamic, see (Ailliot and Monbet, 2012) where a hidden regime-switching model is used for wind speed at one site. In (Pinson et al., 2008), several regime-switching models are compared for the forecast of wind power, hidden MS-AR models are shown to outperform the other proposed models, namely an ARMA model and regime-switching models such as Self-Exiting Threshold AR model (SETAR).

1.4.2 Including multi-site interactions

Recent researches focus on multi-site stochastic weather generator and in addition to the previous requests, one has to try to reproduce the observed dependence between sites. Some meteorological variables exhibit strong correlation that should not be ignored. When considering datasets of multi-site time series, space-time interactions linked to the propagation of meteorological events, should be embedded. Space-time patterns of meteorological data generally lead to properties of non-separability of space and time components of the associated covariance function.

One way of accounting for spatial dependence of multi-station dataset is to work in a multivariate framework, like a Gaussian multivariate framework (Bardossy and Plate, 1992; Yang et al., 2005) or like models based on multivariate ARIMA (Haslett and Raftery, 1989). Random Gaussian fields are well adapted to model space-time interactions through the second order structure of the process (Rychlik and Mustedanagic, 2013; Fuentes et al., 2005). A more general framework than the one of random Gaussian fields is the one where dependences are modeled through copula (Tastu et al., 2013). However this framework leads to more challenging estimation procedures. A wide variety of space-time covariance models that deal with space-time interaction have been proposed in the literature (Gneiting, 2002). Another way of accounting the multi-site dependence is to explicitly simulate the physical phenomena that generate weather (Cox and Isham, 1988).

Another approach is to consider single-site model and to add an extra layer that embeds the spatial distribution of the estimate parameters (Šaltytė Benth

and Šaltytė, 2011) or that gives a spatial structure to random numbers that serve to simulate the variables (Wilks, 1998; Kleiber et al., 2012; Khalili et al., 2007; Thompson et al., 2007). In (Wilks, 1998; Kleiber et al., 2012; Khalili et al., 2007), latent spatial Gaussian processes are introduced to generate the random numbers that drive the precipitation occurrences and amounts. When considering multi-layer models, multi-site dependencies can be assessed in one layer (Bardossy and Plate, 1992) or more, in (Thompson et al., 2007) for instance both layers involve spatial structure.

Space-time propagations of meteorological events can also be introduced using latent variables, in (Ailliot et al., 2006b), the Vector AutoRegressive coefficients depend on a latent process that describes the motion of the air masses.

We work in Chapters 2 and 4 with multivariate processes and models, that account for two space-time scales. In both Chapters 2 and 4, we introduce a regional process that intends to help the description of a part of the propagation of meteorological events. In Chapter 2, we propose a Gaussian linear state-space model, where the local observation is written as a linear projection of the regional wind, this latter is described by a hidden autoregressive scalar process, the projection matrix enables to account for a part of the space-time motions. In Chapter 4, we introduce multivariate Markov-Switching Autoregressive models with observed and latent regional switches. In this modeling, the local process has its own autoregressive dynamic conditionally to the regional scale, which is described by a discrete Markov chain. In this model the space-time interaction is in part included in the autoregressive matrices.

We show that all these proposed models enable to well reproduce a part the distribution of the considered processes and the general shape of space-time covariance functions. To account for space-time variability at another scale than the average scale and reproduce the observed regime-shifts, one can introduce regime-switching models, as described above.

1.4.3 Accounting for regime-switching

Typical patterns of regime-shifts observed on time series can be described by regime-switching models at single or multiple sites. Space-time motions of meteorological events are linked to the current weather state. Indeed storms and stable conditions occur in privileged regional weather conditions. Consequently describing weather state enables to have information about space-time motions of meteorological events. Blocking a time series into regimes consists in partitioning it into periods of time in which the series is homogeneous and can be described by a single process. In that context, weather variables are modeled conditionally to the regime; and the choice of the regimes and of the conditional distribution are of great importance. In most cases, the regime-switching has a Markovian dynamic (Richardson, 1981; Wilks, 1998), but non-

parametric methods are proposed in (Racsko et al., 1991). Regime-switching introduces in the modeling framework various temporal scales. Indeed variations of regimes is at a larger temporal scale than the one of the observations which is a small-scale dynamic. Besides in terms of temporal dependence, regime-switchings enable to combine several dynamics into one model. Indeed it allows to alternate between periods with high temporal variability and more stable periods.

Depending on the availability of good descriptors of the current weather state, regime-switching can be achieved through models with an observed or a latent regime-switching. In the first case, regimes can be extracted via clustering methods from extra-variables, such as descriptors of atmospheric circulation (see for instance (Bardossy and Plate, 1992; Wilson et al., 1992)), or from the studied local variables. A separation of dry-wet states has been widely used to derive observed regimes when various meteorological variables are considered (Richardson, 1981; Flecher et al., 2010). When considering wind models, wind direction can be accounted for since it is a good descriptor of synoptic conditions. In (Gneiting et al., 2006), wind direction is either used to extract regimes or in the parametrization of the predictive distribution. In the second case, when the regimes are not observed, they are then introduced as a hidden variable, see for instance (Ailliot and Monbet, 2012). In that case, the model falls into the domain of Hidden Markov Model. Hidden Markov Model have been widely used in this context for meteorological data (Zucchini and Guttorp, 1991; Hughes et al., 1999; Thompson et al., 2007). In (Ailliot and Monbet, 2012) wind speed at one site is modeled by a hidden MS-AR model.

In the multi-site context, the regime can be regional, common to all sites, and remains scalar (Ailliot et al., 2009) or it can be introduced as a site-specific regime (Wilks, 1998; Kleiber et al., 2012; Khalili et al., 2007; Thompson et al., 2007), which enables to account for a wide range of space-time dependence. However a site-specific regime appears to be computationally challenging (Wilks, 1998).

In Chapters 3 and 4, we use Markov-Switching AutoRegressive models. In Chapter 3, we describe polar and Cartesian coordinates at one site via hidden MS-AR models, in this modeling transitions between the regimes are driven by wind direction as we suggest in Section 1.2. In Chapter 4, Cartesian components at several stations are modeled by Vector Autoregressive models with regime-switching. In this multi-site context, we show that the use of a regional regime is reasonable.

To the best our knowledge, no comparison between the use of observed and latent shifts have been conducted. In this aim, we propose in Chapter 4 several *a priori* regime-switching models and a hidden one and we compare them. The comparison is lead in terms of meteorological consistency of the extracted regimes, of appropriate description of the associated conditional dis-

tribution and of ability to reproduce several chosen statistics. The observed regime-switching models are based on several classifications extracted from a large-scale descriptor of atmospheric circulation and from the local wind conditions. We discuss the difficulty to find physically consistent *a priori* regime that are also appropriate to the description of the conditional model in an autoregressive framework. The hidden regime-switching framework seems the most appropriate to cope with this compromise.

As the meteorology in Europe is driven by large-scale North-Atlantic weather regimes, we investigate the link between the weather state (hidden or observed) extracted from the local wind with these weather regimes. Finally we highlight the benefit of using regime-switching models in the modeling of the alternate of temporal and intensity variabilities in wind conditions.

1.5 Accounting for interaction between wind direction and speed

When considering wind fields one may account for both intensity and direction. We propose in the two last chapters, models that encompass information from both intensity and direction. Moreover, to improve the description of space-time motions of wind events, we seek to model this interaction. Indeed, space-time motions of wind events are associated with characteristic patterns of wind speed and direction. Besides wind direction is a good descriptor of synoptic conditions. The interaction between direction and intensity is very complex. Consequently in a first time, in Chapter 3, we focus on modeling this dependence at a single site through the modeling of polar and Cartesian components, then an extension to the multi-site context is proposed in Chapter 4 for the Cartesian components. This extension is challenging since space-time interaction and the dependence between wind direction and intensity are to be accounted for.

1.5.1 Modeling wind direction time series

Wind direction is of great interest when modeling wind conditions since this variable is very informative about the non-linear behavior of wind speed. In comparison with the literature on the modeling of time series of wind speed, there exists only very few models for time series of wind direction. However, various methodologies have been proposed in the literature to describe the temporal features of circular time series.

In (Breckling, 1989) two autoregressive models for directional data are introduced: von Mises processes and models based on wrapping a linear process on the circle. A generalization of this wrapping is introduced in (Fisher and Lee, 1994) as the linked processes, a Markov model is introduced in (Kato,

2010). As for linear time series, to account for non-linearities of dynamics of circular time series, regime-switching models have been developed (McDonald and Zucchini, 1997). In (Holzmann et al., 2006), various Hidden Markov Models with emission probabilities like von Mises distributions are investigated and applied to two sets of directional data. In (Muñoz et al., 2013), two models with regime-switching and circular probabilities of emission are introduced, the first is a Markov-Switching with von Mises conditional distribution of emission and the second is a model with similar emission probabilities with a deterministic mechanism of thresholds that governs the change of regime.

In Chapter 3, a hidden Markov-switching model is proposed for wind direction time series. The modeling is based on a von Mises process with hidden Markovian shifts between regimes. As described in Section 1.2, transitions between regimes occur in privileged direction, we propose a parameterization of transitions between regimes that depends on wind direction. This leads to an accurate modeling of wind direction time series (marginal distribution, temporal dependence and frequency of clockwise rotations). The comparison with the first model proposed in (Muñoz et al., 2013), where transitions are homogeneous, reveals that the model we suggest gives a better description of wind direction time series.

1.5.2 Modeling polar and Cartesian coordinates of wind

To model the dependence between wind speed and direction, we propose models for polar and Cartesian components of wind. Various aspects of the dependence between wind speed and direction have been explored namely (Qin et al., 2010) gives an example of a joint distribution, in (Modlin et al., 2012) a spatial modeling of polar coordinates of wind fields in hurricanes is proposed. In (Holzmann et al., 2006), a Hidden Markov Model for linear-circular data is proposed but the dependence between the linear and circular variables is ignored. However to the best of our knowledge, no models have been introduced to generate jointly time series of wind speed and wind direction and to account for the dependence between these variables of different nature in a temporal framework.

In Chapter 3, we propose a model to generate simultaneously time series of wind speed and direction and that accounts for a part of the dependence between speed and direction. The proposed model is made of four layers, two of them represent the model for wind direction quoted above, which is a Markov-Switching von Mises process. The two remaining layers are dedicated to describe wind intensity U by a hidden Markov-Switching AutoRegressive model. Transitions between regimes are driven by wind direction. The structure of conditional distributions of the model, with two layers of hidden variables $S^{(\Phi)}$ and $S^{(U)}$, one for the wind speed and one for the wind direction, is shown on

the directed graph below.

$$
\begin{tikzcd}
\cdots \arrow[r] \arrow[d] & S_{t-1}^{(\Phi)} \arrow[r] \arrow[d] & S_{t}^{(\Phi)} \arrow[r] \arrow[d] & S_{t+1}^{(\Phi)} \arrow[r] \arrow[d] & \cdots \arrow[d] \\
\cdots \arrow[r] \arrow[ur] \arrow[dr] & \Phi_{t-1} \arrow[r] \arrow[ur] \arrow[dr] & \Phi_{t} \arrow[r] \arrow[ur] \arrow[dr] & \Phi_{t+1} \arrow[r] \arrow[ur] \arrow[dr] & \cdots \\
\cdots \arrow[r] \arrow[d] & S_{t-1}^{(U)} \arrow[r] \arrow[d] & S_{t}^{(U)} \arrow[r] \arrow[d] & S_{t+1}^{(U)} \arrow[r] \arrow[d] & \cdots \arrow[d] \\
\cdots \arrow[r] & U_{t-1} \arrow[r] & U_{t} \arrow[r] & U_{t+1} \arrow[r] & \cdots
\end{tikzcd}
$$

This model is limited to the single-site context, the extension to a multi-site framework of the model for wind direction is not natural and challenging. Models for Cartesian coordinates are more easily and naturally extended to a multi-site context.

As far as we know, only a few models have been proposed to model time series of Cartesian coordinates of wind (u, v) and the proposed ones are not purposed to wind conditions generation and not focused on reproducing the same statistics we are interested in. In (Hering and Genton, 2010), a VAR model for wind prediction with skew-t-distribution of innovation is proposed for zonal and meridional wind components. In (Ailliot et al., 2006b), Cartesian components of wind are modeled through an autoregressive model with coefficients varying in time according to a hidden state that represents the displacement of the field between two consecutive times. In (Wikle et al., 2001), a Bayesian spatio-temporal framework that accounts for a physical description of large-scale variations is introduced for surface wind fields. In (Fuentes et al., 2005), Gaussian models with high structured non-separable space-time covariance are proposed to model wind fields.

In Chapters 3 and 4, two stochastic generators of (u, v)-conditions are respectively proposed for a single-site dataset and a multi-site one. We propose a hidden Markov-Switching AutoRegressive framework to describe this bivariate process at one location and its extension to multiple locations. In the single-site case, transitions are driven by wind direction, however in the multi-site context we consider homogeneous transitions since the choice of the co-variate is delicate in a multi-site framework. In Chapter 4, we highlight the difficulty to reproduce marginal distribution and temporal patterns of the data contrarily to the Chapter 3, where these statistics are well described by the single-site models for polar and Cartesian components. To the extent of our knowledge, these models constitute one of the first stochastic models that generate jointly time series of Cartesian components.

1.6 Plan of the thesis

In a first step to model multi-site time series of wind, we focused on wind speed and reproducing motions of wind events. In the Chapter 2, we propose a linear Gaussian state-space model as a stochastic generator of wind speed time series at multiple stations. We investigate the identifiability of the model through the study of the second order structure of the model. We propose here two methods to perform efficiently the statistical inference. One is based on the generalized method of moments and the other on maximum likelihood via the EM-algorithm. Various reduced models are also introduced to improve the parsimony of the model. Validation of the model is performed through simulations and one-step ahead prediction is used as a complementary tool for validation.

Limitations of this proposed model are mostly due to the scalar nature of the hidden state, which can not capture all the information of the wind field. Indeed only average behaviors are captured with this model. In order to capture consistent instantaneous behaviors linked with the current weather state and especially the alternate of temporal variability of wind conditions, we consider models with regime-switching. Moreover to account for all the information of wind fields, we added wind direction in the modeling framework through the consideration of polar and Cartesian coordinates of wind.

In Chapter 3, we propose several hidden Markov-switching models for wind condition time series that explicitly model the current weather state. A model for polar components of wind and one for Cartesian components are proposed in a single-site framework. The temporal dynamic of Φ and the joint distribution of (u, v) reveal very complex patterns. We show that the proposed models give a satisfying description of complex features of the studied time series such the joint and marginal distribution, rotations of (u, v) or temporal dynamics.

In Chapter 4, we propose an extension to a multi-site context of the model proposed for (u, v)-components in Chapter 3. We discuss the choice of the regime-switching, which can be observed or latent, and its computation, indeed we compare several classifications extracted from a large-scale descriptor of atmospheric circulation and from the local wind conditions. Moreover we highlight the link between the weather state (hidden or observed) extracted from the local wind with large-scale North-Atlantic weather regimes. We highlight the difficulty to find physically consistent *a priori* regime that are also appropriate to the description of the conditional model in an autoregressive framework. Finally we highlight the benefit of using regime-switching in the modeling of the alternate of temporal variability in wind conditions.

Chapter 2

A multi-site Gaussian linear state-space for wind speed

This chapter is accepted for publication in *Environmetrics*:
Bessac, J., Ailliot, P. and Monbet, V. (2014). Gaussian linear state-space model for wind fields in the North-East Atlantic. *Environmetrics*, to appear

A multi-site stochastic generator for wind speed is proposed. It aims at simulating realistic wind conditions with a focus on reproducing the space-time motions of the meteorological systems. A Gaussian linear state-space model is used where the latent state may be interpreted as regional wind conditions and the observation equation links regional and local scales. Parameter estimation is performed by combining a method of moment and the EM algorithm whose performances are discussed using simulation studies. The model is fitted to 6-hourly reanalysis data in the North-East Atlantic. It is shown that the fitted model is interpretable and provides a good description of important properties of the space-time covariance function of the data, such as the non full-symmetry induced by prevailing flows in this area.

2.1 Introduction

Many natural phenomena and human activities depend on wind conditions. However meteorological data are often available over periods of time that are not long enough to estimate reliably probabilities of complex events. In order to overcome this insufficiency, stochastic weather generators have been developed. Those stochastic weather generators are statistical models that simulate sequences of meteorological variables with statistical properties similar to the ones of the observations. They have been adopted in impact studies as a computationally inexpensive tool that generates quickly as many synthetic time series of unlimited length as desired (see (Srikanthan and McMahon, 1999) and references therein). Stochastic weather generators can be adapted to in-filling

tools that simulate missing data (Yang et al., 2005) or for downscaling global climate models (see e.g. (Maraun et al., 2010) and references therein). Wind generators have in particular been used to assess various quantities related to wind power production (Brown et al., 1984; Castino et al., 1998; Hofmann and Sperstad, 2013), drift of objects in the ocean (Ailliot et al., 2006a) or coastal erosion (Skidmore and Tatarko, 1990).

Many natural phenomena and human activities depend on wind conditions. However meteorological data are often available over periods of time that are not long enough to estimate reliably probabilities of complex events. In order to overcome this insufficiency, stochastic weather generators have been developed. Those stochastic weather generators are statistical models that simulate sequences of meteorological variables with statistical properties similar to the ones of the observations. They have been adopted in impact studies as a computationally inexpensive tool that generates quickly as many synthetic time series of unlimited length as desired, see for instance (Srikanthan and McMahon, 1999) and references therein. Stochastic weather generators can be adapted to in-filling tools that simulate missing data (Yang et al., 2005) or to downscaling global climate models, see for instance (Maraun et al., 2010) and references therein. Wind generators have in particular been used to assess various quantities related to wind power production (Brown et al., 1984; Castino et al., 1998; Hofmann and Sperstad, 2013), drift of objects in the ocean (Ailliot et al., 2006a) or coastal erosion (Skidmore and Tatarko, 1990).

A review of stochastic models for wind time series can be found in (Monbet et al., 2007). Most of the existing models are designed for wind time series at a single location. The most classical approach consists in using the Box-Jenkins methodology, where an ARIMA model is fitted after achieving stationarity and applying a marginal transformation to obtain Gaussian like margins. Non-linear models have also been proposed and, in particular, weather type models with a discrete latent variable, see (Ailliot and Monbet, 2012) and references therein.

Generalizations to space-time models have been explored recently. Multisite wind models have to deal with the temporal and spatial dependence and it is known that these two components are generally not separable when air masses are moving in a prevailing direction (Gneiting, 2002). Black-box models such as artificial neural networks may be fitted but they lead to non-interpretable models (Lei et al., 2009). A first alternative is based on Gaussian fields (Gneiting, 2002; Rychlik and Mustedanagic, 2013) where non-separable parametric covariance functions can be considered to take into account the mean displacement of the air masses. Another approach consists in using vector AutoRegressive-Moving-Average models (Haslett and Raftery, 1989; de Luna and Genton, 2005) where the wind dynamic is described by the autoregressive matrices. Motions can be introduced using covariates or latent variables. For example, in (Ailliot et al., 2006b) the autoregressive coefficients

depend on a latent process that describes the motion of the air masses. In (Šaltytė Benth and Šaltytė, 2011), a latent field describes the spatial structure of the autoregressive parameters at each station. Following similar ideas, various authors developed models that aim at embedding physical insights into a probabilistic model. The Bayesian framework is very convenient to deal with such coupling (Wikle et al., 2001). For instance, in (Milliff et al., 2011), classical partial differential equations for the wind at the sea surface are perturbated by adding a white noise and the parameters are estimated following a Bayesian inference method.

In the present work, a structural model which aims at simulating wind speed at several locations is investigated. The main idea consists in introducing a latent variable which aims at describing regional wind conditions and the observed local wind is modeled as a function of the regional wind at different lags to reproduce the mean displacement of the air masses. The model is kept simple with linear Gaussian models used to describe both the dynamics of the latent process and the link between the latent and the observed process. It leads to an interpretable model with efficient numerical procedures available for parameter estimation and simulation. Despite its simplicity, the model leads to non-separable and anisotropic covariance functions. No physical equations were embedded because their resolution is generally computationally too expensive for a stochastic generator but the suggested model involves quantities that have a physical meaning in the proposed context. It could be used as a surrogate of the atmospheric model (emulator) for data assimilation or data fusion.

The data considered in this work are presented in Section 2.2. The model is described in Section 2.3. Parameter estimation and fitting procedures are also discussed in this section. Validation of the model is discussed in Section 2.4. It is shown that the fitted model is able to reproduce the anisotropy and non-separability of the data. However, the model includes a large number of parameters and various reduced models are introduced in Section 2.5. Conclusions are given in Section 2.6. Parameter identifiability and non full-symmetry are proven in Appendix 2.7.

2.2 The wind dataset

In situ data are neither available on a long time period nor on a large area offshore Brittany in France. Reanalysis data, which are obtained by combining assimilation of observations with numerical weather prediction models, provides a relevant alternative for meteorological or climatological studies. In this work we consider wind speed at 10 meters above sea level extracted from the ERA Interim Full dataset produced by the European Center of Medium-range Weather Forecast (ECMWF). It can be freely downloaded and used for

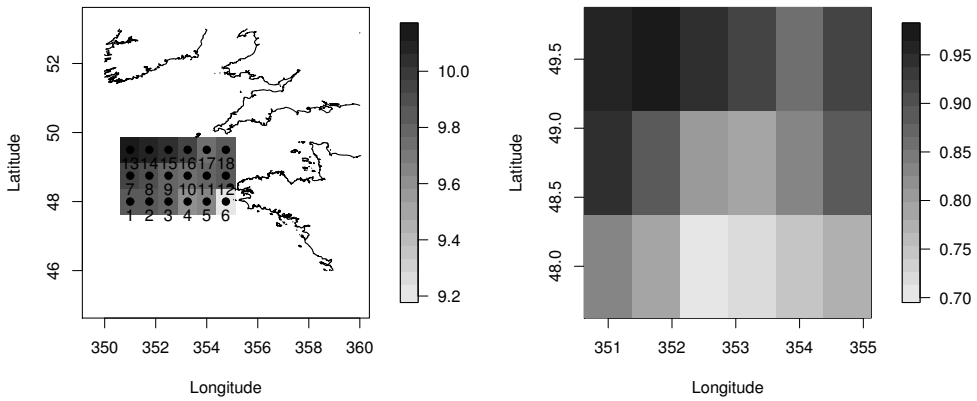

Figure 2.1: Left panel: mean wind speed at the 18 numbered points under study in the North-East Atlantic. Right panel: estimated values of the power in the Box-Cox method at the 18 locations.

scientific purposes at the URL *http://data.ecmwf.int/data/*. This dataset is available on a regular space-time grid with a temporal resolution of 6 hours and a spatial resolution of 0.75°. However the methodology introduced in this work could easily be adapted to handle datasets with more complicated space-time sampling such as the one obtained when considering networks of meteorological stations.

We focus on 18 gridded locations between latitudes 48°N and 49.5°N and longitudes 6.25°W and 9°W (see Figure 4.1). The dataset consists of 33 years of wind data from 1979 to 2011 and we focus on the month of January. Further, the statistical inference is based on the assumption that the 33 months of January wind data are 33 independent realizations of a common stationary stochastic process. This assumption is usual for meteorological processes but it does not take into account low frequency variations such as the North Atlantic Oscillation (NAO).

In the studied area prevailing air masses are generally moving eastward. It induces non-separability and non full-symmetry properties of the space-time covariance function of the wind speed as for the dataset of wind speed in Ireland considered in (Haslett and Raftery, 1989) and (Gneiting, 2002). The lagged by 1 cross-correlations shown in Figure 2.2 highlight this phenomenon. Indeed, the asymmetry with respect to the difference of longitude shows that the correlation between $y_t(p)$ and $y_{t+1}(p')$ is higher when location p is more westerly with respect to p' than when p is easterly with respect to p' and thus that western locations see the meteorological events before the eastern

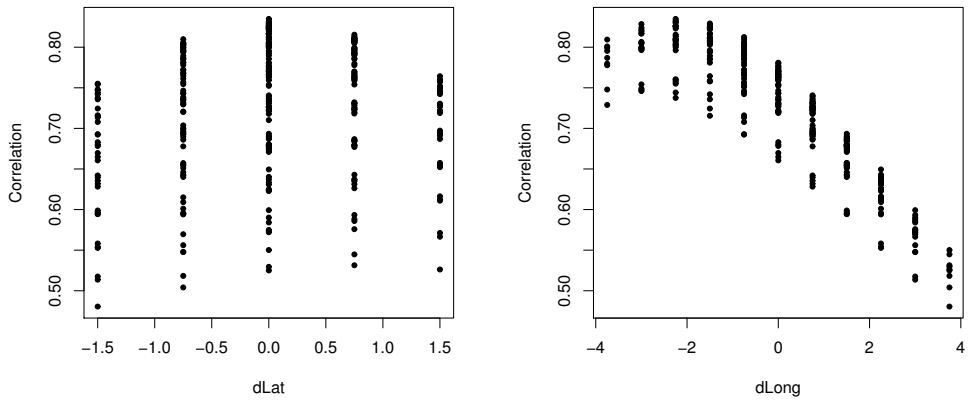

Figure 2.2: Lagged-one cross-correlations against differences of latitude (left) and longitude (right).

locations. This asymmetry is less pronounced in latitude but reveals flows from north to south. Furthermore, the correlations at lag 0 reveals some anisotropy as dependences in latitude and longitude differ (see Figure 2.2).

Wind speed distribution is known to be skewed. It is often modeled as a Weibull distribution (see e.g. (Brown et al., 1984)) but other distributions such as the skew normal distribution have also been considered (see (Flecher et al., 2010)). A classical method to handle such asymmetry in time series analysis consists in applying a Box-Cox transformation in order to get a time series with approximately Gaussian marginal distribution. This method has been extensively used for analyzing wind time series at a single location (see e.g. (Brown et al., 1984)). In (Rychlik and Mustedanagic, 2013) a different power transformation λ_i is used at each location. More precisely, let us denote

$$\begin{cases} y_{\lambda_i,i,t} &= \frac{y_{i,t}^{\lambda_i}-1}{\lambda_i} \quad \text{if } \lambda_i > 0 \\ y_{\lambda_i,i,t} &= \log(y_{i,t}) \quad \text{if } \lambda_i = 0. \end{cases}$$

with $y_{i,t}$ the wind speed at time t and location i. Following (Hinkley, 1977), λ_i can be estimated by searching the roots of the asymmetry measure

$$S(\lambda_i) = \frac{\text{mean}(y_{\lambda_i,i,t}) - \text{median}(y_{\lambda_i,i,t})}{\sqrt{\text{var}(y_{\lambda_i,i,t})}}. \tag{2.1}$$

The resulting estimates are shown in Figure 4.1 with values ranging from about 1 (Gaussian distribution) in the north-west to .7 closer in the south-east. Despite this spatial variability, we have chosen to use the same power

transformation at all sites in order to preserve the spatial structure of the wind fields following e.g. (Haslett and Raftery, 1989). The value $\hat{\lambda} = 0.85$ is used in the sequel. It is the average value of the $\hat{\lambda}_i$ shown on Figure 4.1. The simulation results given in Section 2.4 (see e.g. Figure 2.5) indicate that this simple transformation permits to reproduce the marginal distributions of the wind data considered in this study.

2.3 A linear Gaussian state-space model for wind speed

State-space models first appeared in engineering and have then been extensively used in many domains. State-space representations bring a very flexible framework for modeling time series (see (Durbin and Koopman, 2012) and (Brockwell and Davis, 2006)) and space-time processes (see (Wikle and Hooten, 2010)). The model introduced in this section is a linear Gaussian state-space model. One of the main advantages of this class of models is that estimation, forecasting and smoothing can be processed through general and efficient procedures.

2.3.1 Model

The observed wind fields are generally smooth, which leads to a high correlation between the different sites. Although the smoothness observed here is inherent to reanalysis data that are known to be smoother than observations (see (Milliff et al., 2011)), it is coherent with the considered spatio-temporal scale. This regularity suggests to explain an important part of the multisite wind by using a common scalar process (the 'regional wind condition'). This scalar process, denoted by $\{X_t\}$ in the sequel, can not be observed directly and is thus introduced as a latent (or 'hidden') process. In order to model the prevailing motion of the air masses we propose to let the wind conditions at western locations depend more on the leading one-lag X_{t+1} and X_t signals than on the lagged signal X_{t-1} with the reverse phenomenon at eastern locations. More precisely, the Gaussian state-space model which is considered in this work is defined as small-scale fluctuations. In finance and economics this covariance matrix of error of measurement is often assumed to be diagonal. Here it would imply that the local wind conditions are conditionally independent given the regional conditions which is a very strong assumption. As a first step we have chosen to work with a full non-parametric covariance matrix but reduced parametric models are explored in Subsection 2.5.1 (see also (Wikle and Hooten, 2010)). In the sequel, we denote $\mathbf{\Lambda} = (\boldsymbol{\alpha}_1|\boldsymbol{\alpha}_0|\boldsymbol{\alpha}_{-1}) \in \mathbb{R}^{K\times 3}$ and $\theta = (\rho, \sigma, \mathbf{\Lambda}, \mathbf{\Gamma})$ the unknown parameters.

The temporal dynamics of the observed process is mainly contained in

the latent process $\{X_t\}$ and explained by the coefficient ρ. The model thus imposes the same long-term temporal dynamics at each location. Under the assumption $|\rho| < 1$, the AR(1) process $\{X_t\}$ is stationary and so is the process $\{\boldsymbol{Y_t}\}$.

2.3.2 Second-order structure and identifiability

Identifiability is required to get sensible and reliable parameter estimates. The introduction of a latent process $\{X_t\}$ is a source of non-identifiability since the unknown parameters need to be identified uniquely from the distribution of the observed $\{\boldsymbol{Y_t}\}$ and Gaussian linear state-space models are known to be often non-identifiable without additional constraints (see e.g. (Hannan and Deistler, 1988), (Ljung, 1999), (Bai and Wang, 2012), (Bork, 2010)). Identifiability of linear Gaussian state-space models has been initially investigated in control theory and was largely explored during the last decades. Literature is abundant on stochastic linear systems identification (Ljung, 1999; Hannan and Deistler, 1988). To the best of our knowledge most of sufficient conditions of identifiability are structural constraints on parameters (see (Papadopoulos and Digalakis, 2010) for references and examples) associated with identification procedures. In most cases structural constraints are applied depending on the interpretability wished. Identifiability is examined through different criteria based on transfer functions (Ljung, 1999) or likelihood (Papadopoulos and Digalakis, 2010). Identification procedures are performed through controllability and observability of several parameters in (Ljung, 1999) and via the EM algorithm in (Papadopoulos and Digalakis, 2010). In econometrics the identifiability of the latent factors and the loading matrix is considered (see for example (Bai and Wang, 2012; Bork, 2010)). However the general conditions given in (Bai and Wang, 2012) do not ensure identifiability of the model (M) since X is scalar in (M). However we could not find any result which applies directly to the model considered in this work.

$\{\boldsymbol{Y_t}\}$ is a zero-mean stationary Gaussian process which is thus characterized by its second-order structure given below

$$\begin{aligned}
\mathrm{cov}_\theta(\boldsymbol{Y}_t, \boldsymbol{Y}_t) &= \frac{\sigma^2}{1-\rho^2}\Big(\boldsymbol{\alpha}_1(\boldsymbol{\alpha}_1 + \rho\boldsymbol{\alpha}_0 + \rho^2\boldsymbol{\alpha}_{-1})^t + \boldsymbol{\alpha}_0(\rho\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_0 + \rho\boldsymbol{\alpha}_{-1})^t + \\
&\quad \boldsymbol{\alpha}_{-1}(\rho^2\boldsymbol{\alpha}_1 + \rho\boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_{-1})^t\Big) + \boldsymbol{\Gamma}, \qquad (2.2)\\
\mathrm{cov}_\theta(\boldsymbol{Y}_t, \boldsymbol{Y}_{t+1}) &= \frac{\sigma^2}{1-\rho^2}\Big(\boldsymbol{\alpha}_1(\rho\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_0 + \rho\boldsymbol{\alpha}_{-1})^t + \boldsymbol{\alpha}_0(\rho^2\boldsymbol{\alpha}_1 + \rho\boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_{-1})^t + \\
&\quad \rho\boldsymbol{\alpha}_{-1}(\rho^2\boldsymbol{\alpha}_1 + \rho\boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_{-1})^t\Big), \qquad (2.3)\\
\mathrm{cov}_\theta(\boldsymbol{Y}_t, \boldsymbol{Y}_{t+k}) &= \frac{\sigma^2}{1-\rho^2}\rho^{k-2}(\boldsymbol{\alpha}_1 + \rho\boldsymbol{\alpha}_0 + \rho^2\boldsymbol{\alpha}_{-1})(\rho^2\boldsymbol{\alpha}_1 + \rho\boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_{-1})^t, \qquad (2.4)\\
&\quad \text{for all } k \geq 2.
\end{aligned}$$

The study of this space-time covariance function leads to the following Proposition which is proven in Appendix 2.7.

Proposition 1 *Assume that* (M) *holds. Assume further that* $\frac{\sigma^2}{1-\rho^2} = 1$ *and that the vectors* $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_0$ *and* $\boldsymbol{\alpha}_{-1}$ *are linearly independent. Then the parameters can be identified from the distribution of the process* $\{\boldsymbol{Y_t}\}$.

These identifiability constraints are interpretable and were always satisfied when fitting the model to the data. The first one implies that X_t has a unit variance, the variance of the wind at the different locations being explained by the scaling matrix $\boldsymbol{\Lambda}$. The second one implies that Y_t actually depends on the three lagged values X_{t-1}, X_t and X_{t+1} and not only on one or two lagged values.

We will see in Section 2.4 that the proposed model enables to reproduce various complex properties of the observed space-time covariance. Under constraints of Proposition 1, the covariance defined by (2.2-2.4) is neither full-symmetric nor separable (see Appendix 2.7). Other non-symmetric space-time covariance models have been proposed in the literature. Some of them have been fitted to the Irish wind dataset (see for instance (Gneiting, 2002)). They generally rely on strong assumptions such as spatial stationarity and isotropy which are not realistic for our dataset. A noticeable exception is the model proposed in (de Luna and Genton, 2005) which is based on the specification of a vector autoregressive process and captures a part of the anisotropy that is observed on the Irish dataset.

2.3.3 Parameter estimation

Two methods of estimation have been implemented and compared. The first one is a method of moment based on the second-order structure of the process $\{\boldsymbol{Y_t}\}$ given by (2.2-2.4). It consists in numerically minimizing the following objective function

$$\begin{aligned}\theta \;\rightarrow\; & \|\widehat{\mathrm{cov}}(\boldsymbol{Y}_t,\boldsymbol{Y}_t) - \mathrm{cov}_\theta(\boldsymbol{Y}_t,\boldsymbol{Y}_t)\|_2^2 + \|\widehat{\mathrm{cov}}(\boldsymbol{Y}_t,\boldsymbol{Y}_{t+1}) - \mathrm{cov}_\theta(\boldsymbol{Y}_t,\boldsymbol{Y}_{t+1})\|_2^2 \\ & + \|\widehat{\mathrm{cov}}(\boldsymbol{Y}_t,\boldsymbol{Y}_{t+2}) - \mathrm{cov}_\theta(\boldsymbol{Y}_t,\boldsymbol{Y}_{t+2})\|_2^2 + \|\widehat{\mathrm{cov}}(\boldsymbol{Y}_t,\boldsymbol{Y}_{t+3}) - \mathrm{cov}_\theta(\boldsymbol{Y}_t,\boldsymbol{Y}_{t+3})\|_2^2,\end{aligned} \tag{2.5}$$

where $\widehat{\mathrm{cov}}$ denotes the empirical covariance function and $\|.\|_2$ stands for the matrix Frobenius norm. This method, denoted GMM for Generalized Method of Moment in the sequel, is standard in geostatistics (see e.g. (Cressie, 1991)). We have chosen to consider only the first four lags of the autocovariance function when building the objective function (2.5). It corresponds to the minimal number of terms needed to identify the parameters (see Appendix 2.7). Simulation results indicate that including more lags does not lead to more accurate estimates.

The second method performs Maximum Likelihood (ML) estimation using the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). EM algorithm aims at maximizing the incomplete log-likelihood function

$$\theta \to \mathrm{E}(\log(p(X_1, ..., X_T, \boldsymbol{Y}_1, ..., \boldsymbol{Y}_T; \theta))|\boldsymbol{Y}_1^T = y_1^T)$$

by performing recursively two steps (E-step and M-step). For linear Gaussian state-space models efficient numerical procedures exist for both steps. In the E-step, the Kalman recursions lead to an exact computation of the various conditional expectations involved and in the M-step, analytical expressions of the maximizers of the intermediate function are available. More details about the Kalman recursions and EM-algorithm can be found in the supplementary materials.

Both methods are sensitive to the initial parameter value which needs to be chosen carefully. We used the following procedure which involves the properties of the second-order structure of $\{\boldsymbol{Y_t}\}$:

- $\rho = \dfrac{\mathrm{cov}(\boldsymbol{Y}_t, \boldsymbol{Y}_{t+3})_{i,j}}{\mathrm{cov}(\boldsymbol{Y}_t, \boldsymbol{Y}_{t+2})_{i,j}}$ for all $i, j \in \{1, ..., K\}$ is initialized as the empirical mean of $\dfrac{\widehat{\mathrm{cov}}(\boldsymbol{Y}_t, \boldsymbol{Y}_{t+3})_{i,j}}{\widehat{\mathrm{cov}}(\boldsymbol{Y}_t, \boldsymbol{Y}_{t+2})_{i,j}}$.

- $\boldsymbol{\Lambda}$ is estimated by minimizing

$$\begin{aligned}\theta_{\boldsymbol{\Lambda}} \quad \to \quad & \|\widehat{\mathrm{cov}}(\boldsymbol{Y}_t, \boldsymbol{Y}_{t+1}) - \mathrm{cov}_\theta(\boldsymbol{Y}_t, \boldsymbol{Y}_{t+1})\|_2^2 \\ & + \|\widehat{\mathrm{cov}}(\boldsymbol{Y}_t, \boldsymbol{Y}_{t+2}) - \mathrm{cov}_\theta(\boldsymbol{Y}_t, \boldsymbol{Y}_{t+2})\|_2^2\end{aligned}$$

  as a function of $\boldsymbol{\Lambda}$ with ρ being fixed to the value obtained in the previous step. Note that this function does not depend on $\boldsymbol{\Gamma}$ according to (2.3) and (2.4).

- $\boldsymbol{\Gamma}$ is determined by minimizing

$$\theta_{\boldsymbol{\Gamma}} \quad \to \quad \|\widehat{\mathrm{cov}}(\boldsymbol{Y}_t, \boldsymbol{Y}_t) - \mathrm{cov}_\theta(\boldsymbol{Y}_t, \boldsymbol{Y}_t)\|_2^2$$

  as a function of $\boldsymbol{\Gamma}$ with ρ and $\boldsymbol{\Lambda}$ being fixed to the value obtained in the previous steps.

These rough estimates are used as initial conditions of the numerical optimization of the function (2.5) to compute the GMM estimates which in turn are used to initialize the EM algorithm. An extra step could be added to refine the output of the EM algorithm with a numerical optimization of the likelihood function which is known to be more efficient close to local maxima (see (Durbin and Koopman, 2012)). However we did not find any improvements in practice with such a procedure.

	Bias		Sd		RMSE	
Parameters	GMM	ML	GMM	ML	GMM	ML
ρ	0.036	0.004	0.022	0.017	0.042	0.017
$\boldsymbol{\alpha}_1$	[-0.11;-0.009]	[-0.069;-0.019]	[0.065;0.108]	[0.071;0.097]	[0.067;0.149]	[0.068;0.127]
$\boldsymbol{\alpha}_0$	[-0.047;-0.234]	[0.054;0.144]	[0.11;0.182]	[0.11;0.144]	[0.125;0.292]	[0.127;0.228]
$\boldsymbol{\alpha}_{-1}$	[-0.080;0.022]	[-0.035;0.012]	[0.078;0.114]	[0.062;0.104]	[0.086;0.139]	[0.079;0.117]
$\boldsymbol{\Gamma}$	[-0.199;0.007]	[-0.108;0.013]	[0.058;0.367]	[0.029;0.368]	[0.053;0.199]	[0.053;0.115]

Table 2.1: Bias, standard deviation and RMSE of parameters estimates. For the multidimensional parameters, minimal and maximal values are given in brackets.

2.3.4 Properties of the estimates

Under suitable conditions, GMM (see (Newey and McFadden, 1994)) and ML (see (Newey and McFadden, 1994; Shumway and Stoffer, 2006; Hannan and Deistler, 1988; Caines, 1988)) estimators are consistent and asymptotically Gaussian. In order to assess the performance of the estimators for the practical application considered in this work, we perform a simulation study. $N = 100$ independent sets of the size of the studied data are simulated for the parameters set estimated by ML on the wind data. Table 2.1 gives the bias, standard deviation and Root Mean Square Error (RMSE) of ML and GMM estimates computed from the simulations. Bias and standard deviations are low. ML generally outperforms GMM except when estimating $\boldsymbol{\Gamma}$ where both methods give comparable results. Both methods estimate more accurately $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_{-1}$ than $\boldsymbol{\alpha}_0$ and $\boldsymbol{\Gamma}$ is the less accurately estimated quantity.

2.4 Results

In order to validate the proposed model we check its physical realism and its ability to generate wind conditions similar to the ones of the dataset. We compare GMM and ML estimates through this validation in order to investigate their robustness in a practical context.

2.4.1 Interpretability

The loading matrix $\boldsymbol{\Lambda}$ links the latent process to observed wind conditions. The values of $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_{-1}$ shown on Figure 2.3 reveal the site-dependent relations with the latent process. Western locations depend more on X_{t+1} than on X_{t-1} and the reverse is true for eastern locations. This was expected since western locations are the first locations affected when meteorological events enter in

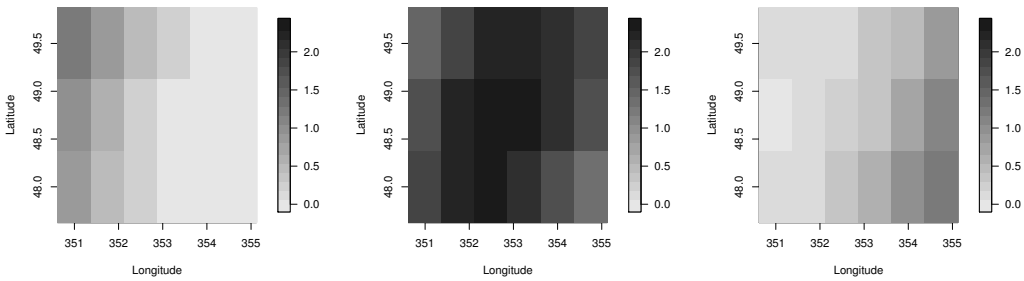

Figure 2.3: ML estimate of $\boldsymbol{\alpha}_1$ (left panel) $\boldsymbol{\alpha}_0$ (middle panel) and $\boldsymbol{\alpha}_{-1}$ (right panel).

the studied region.

Since large-scale variability is supposed to be contained in the latent process, $\boldsymbol{\Gamma}$ should contain only small-scale variations. This is confirmed when comparing the spatial sill and range of $\boldsymbol{\Gamma}$ with the ones of the original covariance function of the data (see Figure 2.4). The shape of $\boldsymbol{\Gamma}$ has a block structure which is induced by the geometry of the domain and the numbering of the sites (see Figure 4.1). The level sets of the blocks, except the top right corner (and by symmetry bottom left corner), look like saddle point level sets: the model better explains the wind observed at the central locations of the domain than at the locations which are close to the boundary. The top right corner has elliptical level sets. These geometrical differences raise problems when trying to develop simple parametric models for $\boldsymbol{\Gamma}$ (see Section 2.5.1).

2.4.2 Realism of simulated sequences

In order to further validate the model, we have checked its ability to simulate realistic wind conditions. For that, artificial time series are simulated with the fitted models and their statistics are compared with the ones of the original data. According to quantile-quantile plots shown on Figure 2.5, the model is able to reproduce the general shape of the marginal distribution of the process at the central station 9 except for very low wind speed. Similar results were obtained at other locations.

Figure 2.7 shows that the cross-correlations at lags 0 and 1 are well reproduced by the fitted models with a slightly better fit for the GMM estimates. This was not unexpected since the GMM estimate is designed to make the first lags of the empirical autocovariance functions coincide with the one of the fitted model. Figure 2.6 shows however that the fit is better for lags greater than

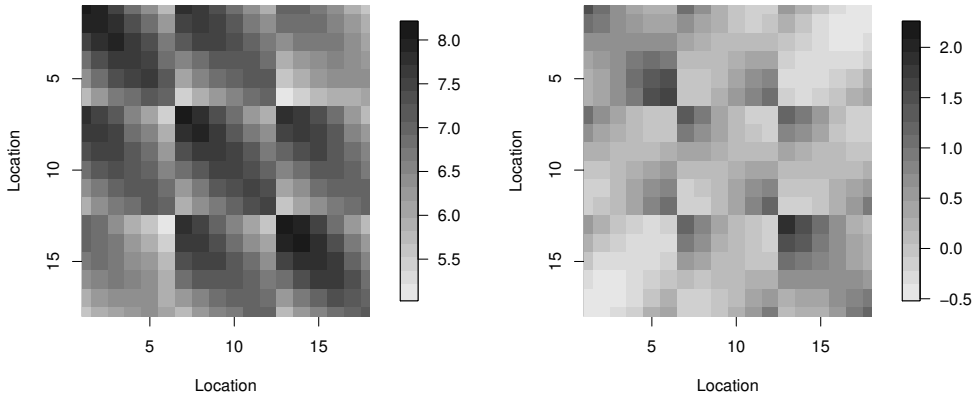

Figure 2.4: Empirical covariance matrix of the wind data (left) and ML estimate of $\mathbf{\Gamma}$ (right).

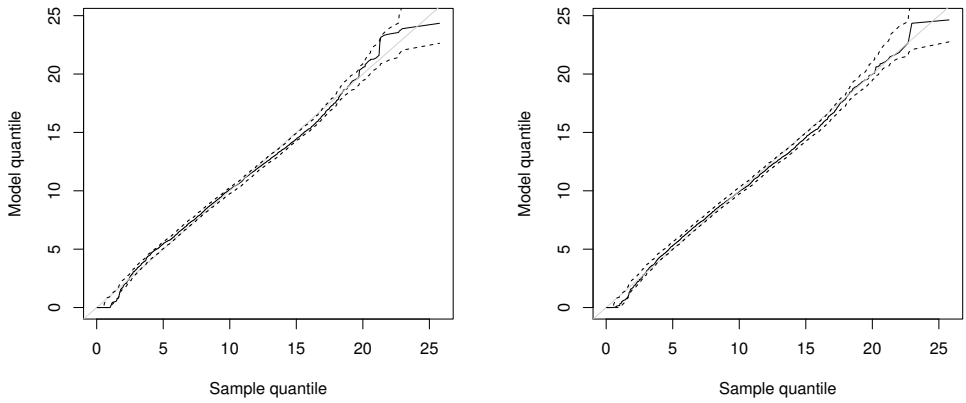

Figure 2.5: Quantile-Quantile plot at location 9 for the model (M) and the parameters estimated by GMM (left) and by ML (right). The dashed lines corresponds to 90% prediction intervals computed by simulation.

one day with the ML estimates which take into account longer term dynamics. It leads to a higher value for ρ (0.76 for ML against 0.70 for GMM). The better fit of the ML estimates is also coherent with with Table 2.1. Note also that the models reproduce the time shift between locations 13 and 18 which is induced by the prevailing westerly flow (see Figure 2.6).

2.5 Some improvements of the model

In this section we explore reduced models for $\mathbf{\Gamma}$ and $\mathbf{\Lambda}$ with the aim of reducing the number of parameters involved in the model.

2.5.1 Parameterization of Γ

The spatial structure of the estimated $\mathbf{\Gamma}$ shown on Figure 2.4 suggests to model the covariance between locations i and j in $\{1, ..., K\}$ as a function of the distance $\mathrm{d}_{i,j}$ between these locations. In the sequel, we consider two different models, one with Gaussian correlation function

$$\mathbf{\Gamma}_{i,j} = \sigma_i \sigma_j (\exp(-\lambda_1 \mathrm{d}_{i,j}^2) + \lambda_2 \delta_{i,j}) \text{ for } i, j \in \{1, ..., K\},$$

and the other with wave correlation function

$$\mathbf{\Gamma}_{i,j} = \sigma_i \sigma_j \Big(\frac{\sin(\lambda_1 \mathrm{d}_{i,j})}{\lambda_1 \mathrm{d}_{i,j}} + \lambda_2 \delta_{i,j} \Big) \text{ for } i, j \in \{1, ..., K\},$$

where $(\sigma_1, ..., \sigma_K, \lambda_1, \lambda_2)$ are positive parameters and $\delta_{i,j}$ denotes the Kronecker delta. λ_1 and λ_2 are respectively the range and nugget parameters, and $\sigma_i^2(1+\lambda_2)$ represents the variance of the field at location i. These models are usually well defined covariance functions (see e.g. (Cressie, 1991; Abrahamsen, 1997)). They are denoted respectively ($\mathrm{M}_{\mathbf{\Gamma} \sim Gauss}$) and ($\mathrm{M}_{\mathbf{\Gamma} \sim Sinus}$) hereafter.

The difference in dependence on latitude and longitude of $\mathbf{\Gamma}$ (Figure 2.8) suggests the use of an anisotropic distance (see (Refice et al., 2011; Haskard, 2007; Šaltytė Benth and Šaltytė, 2011))

$$\mathrm{d}_{i,j} = \sqrt{\Delta\mathrm{Lat}(i,j)^2 + \theta_1 \Delta\mathrm{Long}(i,j)^2 + \theta_2 \Delta\mathrm{Lat}(i,j)\Delta\mathrm{Long}(i,j)}$$

where $\Delta\mathrm{Lat}(i,j)$ and $\Delta\mathrm{Long}(i,j)$ denote respectively the difference in latitude and longitude between locations i and j expressed in kilometers. The constraint $\theta_1 > \frac{\theta_2^2}{4}$ is imposed to ensure the positive-definiteness of the distance.

These covariance structures have first been fitted by least square estimation to the estimated $\mathbf{\Gamma}$ shown on Figure 2.4. Results are shown on the bottom panels of Figure 2.8 and have to be compared with the right-hand side

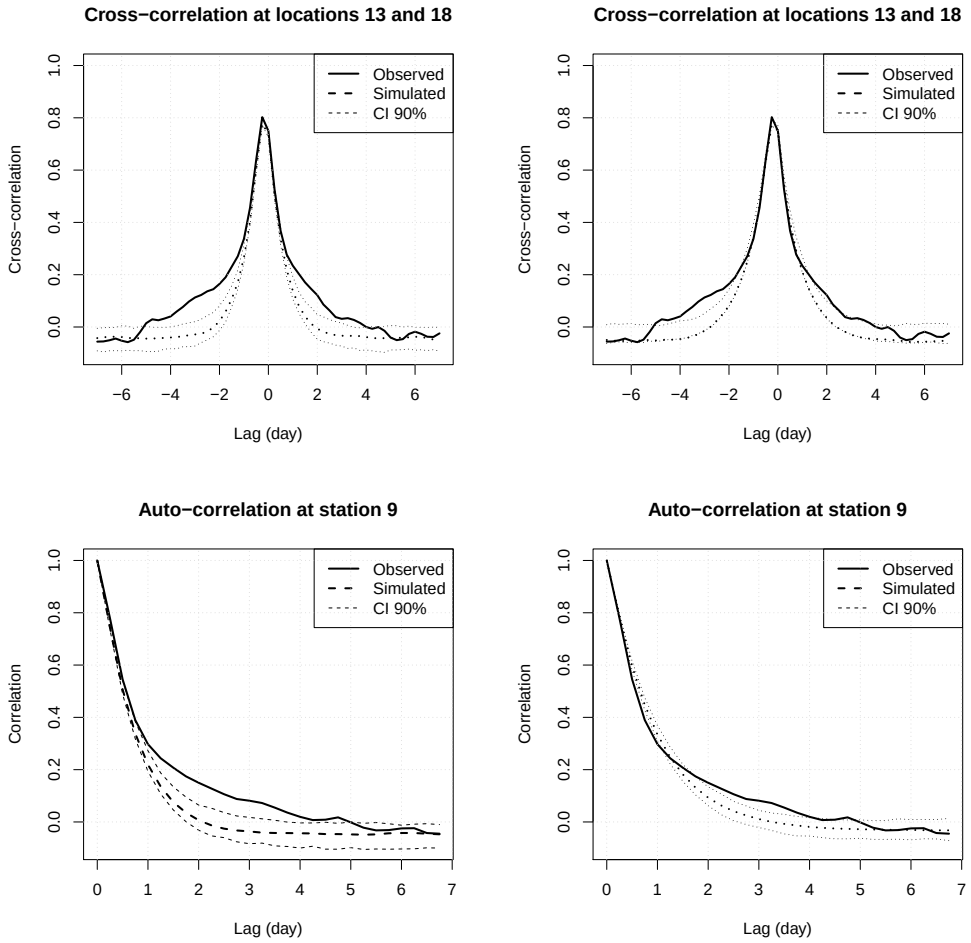

Figure 2.6: Observed (full lines) and theoretical (dashed lines) cross-correlations between locations 13 and 18 (upper row) and auto-correlation at location 9 (lower row) for the model (M) with parameters estimated by GMM (left) and by ML (right). 90% prediction intervals are computed from 100 independent simulated samples of the size of the original data.

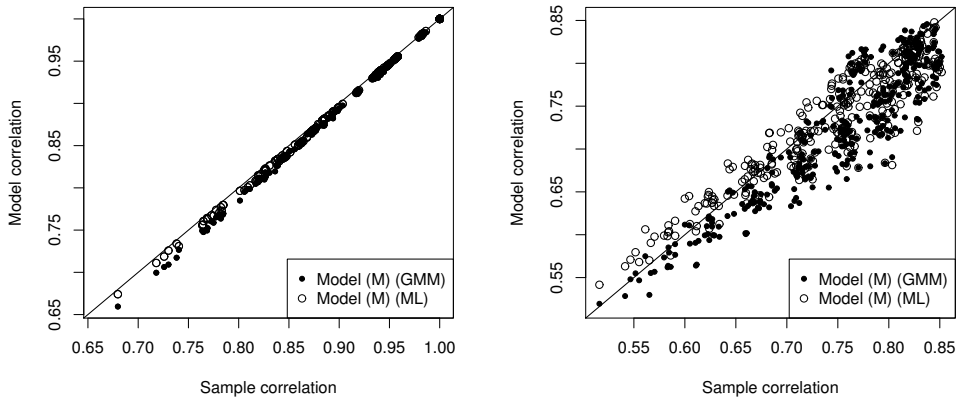

Figure 2.7: Theoretical correlations against observed correlations at lag 0 (left) and lag 1 (right) for the model (M) and the two methods of estimation.

of Figure 2.4. The fit is globally good for the wave covariance whereas the Gaussian shape can not cope with the negative correlations observed between western and eastern locations. However the covariance between the northern and southern locations are poorly reproduced (bottom left corner and top right corner of the images of $\mathbf{\Gamma}$ in Figures 2.8 and 2.4). As mentioned in Section 2.4.1 these blocks have a particular elliptical shape which is difficult to reproduce by parametric models. Estimated anisotropy coefficients for the sinus and the Gaussian structures are respectively $(\hat{\theta}_1, \hat{\theta}_2) = (0.2, 0.04)$ and $(\hat{\theta}_1, \hat{\theta}_2) = (0.23, 0.005)$. For both models θ_1 is lower than one and θ_2 is close to zeros and thus the spatial range is maximum in the west-east direction (see Figure 2.8).

In a second step, the parameters have been re-estimated using the GMM and ML methods. A numerical optimization needs to be performed in the M-step of the EM algorithm to update the values of $(\sigma_1, ..., \sigma_K, \lambda_1, \lambda_2)$. Note that the function to minimize can be expressed in a compact way (see supplementary materials) which leads to an efficient numerical procedure. The models have been validated in the same way as model (M) (see Section 2.4). Similar results were obtained for the marginal distributions and the temporal correlation functions. However the description of the spatial structure was deteriorated when using a ($\mathrm{M}_{\mathbf{\Gamma}}$) model instead of (M) (compare Figure 2.7 with Figure 2.9). This miss-specification is also confirmed by the Bayes Information Criterion (BIC) values given in Table 2.2 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 reduced models ($\mathrm{M}_{\mathbf{\Gamma}}$) are clearly outperformed

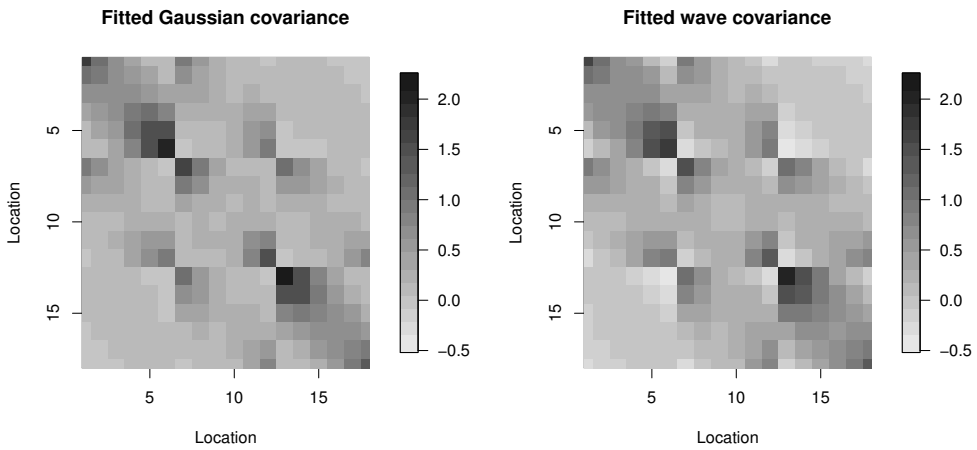

Figure 2.8: Image of covariance matrices fitted by least square to the matrix shown on Figure 2.4 (Gaussian covariance (left), wave covariance (right)).

Model	Parameters	Log-likelihood	BIC
(M_2)	209	-24849	52040
(M)	208	-24954	52238
($M_{\mathbf{\Lambda}}$)	186	-25399	52895
($M_{\mathbf{\Gamma}\sim Gauss}$)	78	-29110	59082
($M_{\mathbf{\Gamma}\sim Sinus}$)	78	-35615	72094

Table 2.2: Table of log-likelihoods and BIC indexes for the different models.

by the full model (M). Other parametric models such as the Matérn one have been tried without more success and it seems difficult to find a simple reduced model which can reproduce all the complexity of the observation error $\mathbf{\Gamma}$.

2.5.2 Parameterization of Λ

The structure of $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_{-1}$ reveals a quadratic dependence in longitude and the dependence in latitude suggests the use of an intercept depending on latitude (see Figure 2.10). This following parameterization is then proposed

$$\mathbf{\Lambda} = \begin{pmatrix} 1 & | & \text{Long} & | & \text{Long}^2 \end{pmatrix} \begin{pmatrix} \beta_1^{\text{Lat}} & \beta_4^{\text{Lat}} & \beta_7^{\text{Lat}} \\ \beta_2 & \beta_5 & \beta_8 \\ \beta_3 & \beta_6 & \beta_9 \end{pmatrix}$$

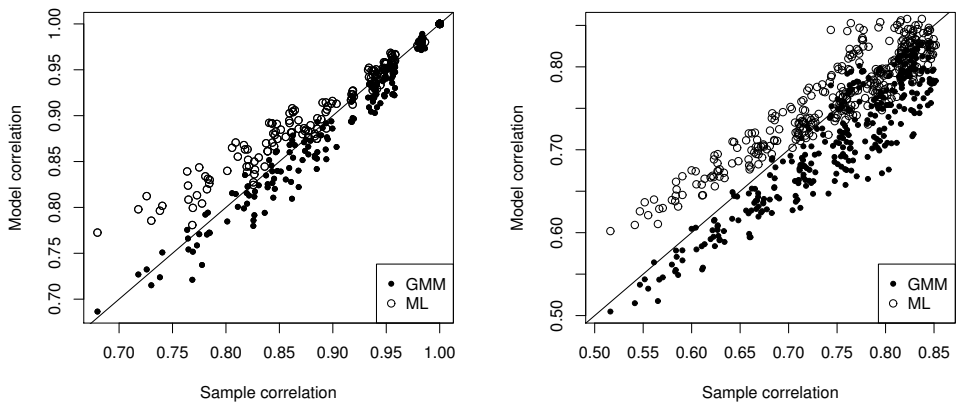

Figure 2.9: Theoretical correlations of the process y for model ($\mathrm{M}_{\mathbf{\Gamma}\sim Gauss}$) against observed ones at lag 0 (left) and lag 1 (right).

where β_i^{Lat} for $i \in \{1, 4, 7\}$ takes a different value for each latitude and $\mathrm{Long} \in \mathbb{R}^K$ is a vector containing the longitude of each site. Let ($\mathrm{M}_{\mathbf{\Lambda}}$) denote the corresponding model. $\mathbf{\Lambda}$ is of rank 3 if the matrix $\begin{pmatrix} \beta_1^{\mathrm{Lat}} & \beta_4^{\mathrm{Lat}} & \beta_7^{\mathrm{Lat}} \\ \beta_2 & \beta_5 & \beta_8 \\ \beta_3 & \beta_6 & \beta_9 \end{pmatrix}$ is full ranked because the matrix $\begin{pmatrix} 1 & | & \mathrm{Long} & | & \mathrm{Long}^2 \end{pmatrix}$ is full ranked.

The parameterization is easily handled in the GMM procedure whereas a numerical optimization is again needed to update $\mathbf{\Lambda}$ in the M-step of the ML procedure. Moreover a joint optimization on $\mathbf{\Lambda}$ and $\mathbf{\Gamma}$ should be done since both of them are involved in the same part of log-likelihood. In order to avoid a numerical optimization in a high-dimensional space, separate optimizations in $\mathbf{\Lambda}$ and in $\mathbf{\Gamma}$ have been performed leading to a so-called Generalized EM algorithm (see the supplementary materials for more details). The reduced ($\mathrm{M}_{\mathbf{\Lambda}}$) and the full model (M) give similar results for the marginal distribution and the autocorrelation function. ($\mathrm{M}_{\mathbf{\Lambda}}$) leads also to an accurate description of the spatial structure of the data (see Figure 2.11). Lagged-one correlations are better reproduced by GMM parameters than by ML parameters. The model ($\mathrm{M}_{\mathbf{\Lambda}}$) is slightly inferior to the full model (M) in terms of BIC according to Table 2.2. Nevertheless, it clearly outperforms the models ($\mathrm{M}_{\mathbf{\Gamma}}$). It seems easier to find an appropriate reduced model for the loading matrix $\mathbf{\Lambda}$ than for the covariance matrix of the observation error $\mathbf{\Gamma}$.

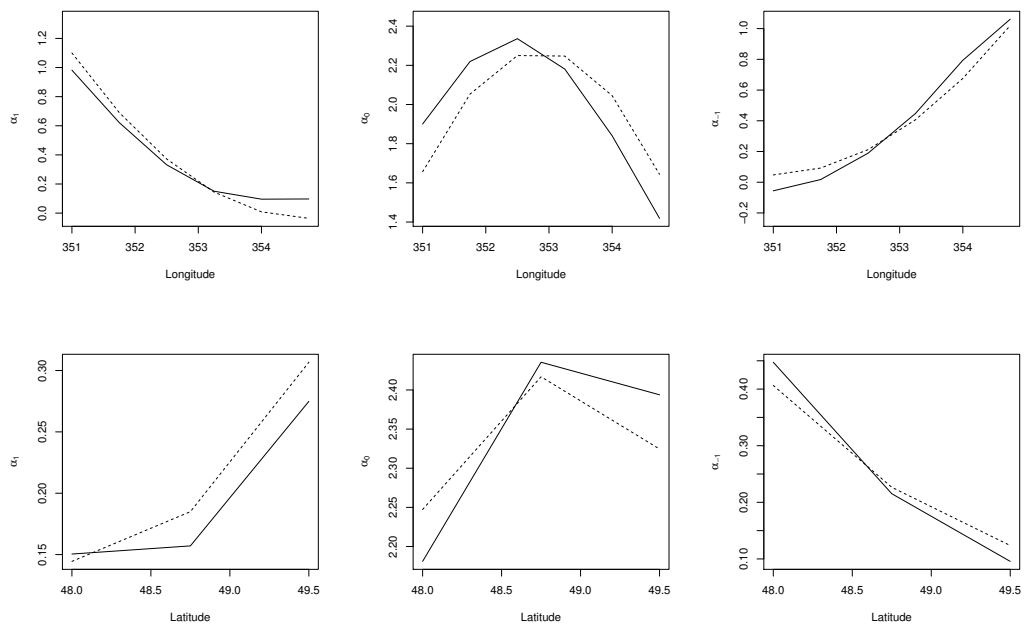

Figure 2.10: Estimated $\boldsymbol{\alpha}_1$ (left), $\boldsymbol{\alpha}_0$ (middle) and $\boldsymbol{\alpha}_{-1}$ (right) against longitude at latitude 48° N (bottom) and against latitude at longitude 6.75° W (top). Solid line: ML estimation of $\boldsymbol{\Lambda}$ for model (M) , dashed line: parametric structure fitted by least square.

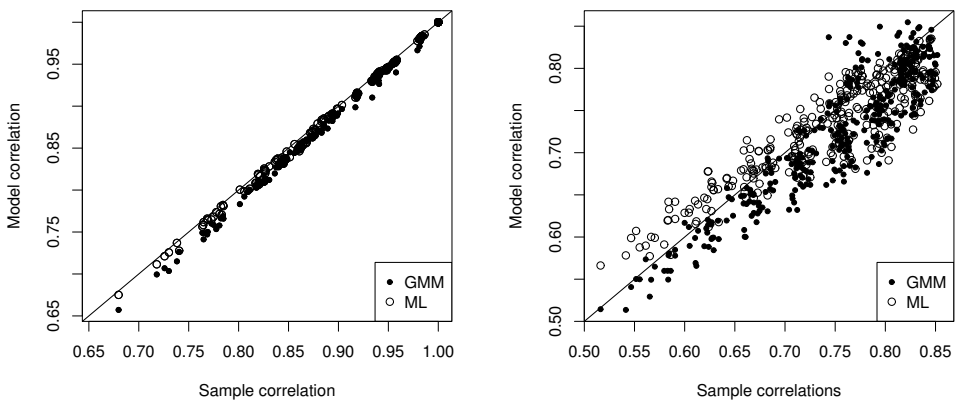

Figure 2.11: Theoretical correlations against observed ones at lag 0 (left) and lag 1 (right) for the model ($\mathrm{M}_{\boldsymbol{\Lambda}}$).

2.6 General discussion

Several multisite models, all based on Gaussian linear state-space models, are proposed to generate synthetic multivariate time series of wind speed. The main innovation, with respect to the other space-time models which have been proposed for meteorological variables, is the introduction of a continuous latent process describing regional conditions. The proposed models are interpretable and can reproduce the marginal distribution of wind speed and important properties of the space-time covariance structure such as the asymmetries induced by prevailing motions of the air masses.

An important advantage of Gaussian linear state-space models is that efficient and easy to implement procedures of estimation are available. Two estimation procedures, one based on a method of moment (GMM) and the other on the likelihood function (ML) have been compared. GMM yields to better results when looking at the short-term space-time structure but ML is better in reproducing the long-term dynamics. Note that higher-order autoregressive models have been considered for modeling the dynamics of the hidden state but they led to very slight improvements and are not further discussed here (the model with autoregressive models of order 2, denoted (M_2), is given in Table 2.2).

According to the BIC values given in Table 2.2 the ranking of the model coincides with the complexity of the model and the quality of the model is systematically worsened when the number of parameters is reduced. In order to check the relevance of the BIC criterion, we have performed a cross-validation study (see supplementary materials) which confirmed the ranking of the models given by BIC. Similar results were obtained on the Irish wind dataset considered in (Haslett and Raftery, 1989; Gneiting, 2002) which has a different space-time resolution with daily data and stations on an irregular spatial grid. This highlights the difficulty to find parsimonious and realistic models for describing the space-time evolution of wind.

2.7 Proof of proposition 1

Let $\{\boldsymbol{Y_t}\}$ [resp. $\{\tilde{\boldsymbol{Y_t}}\}$] denote a process satisfying (M) with parameters $\theta = (\rho, \sigma, \boldsymbol{\Lambda}, \boldsymbol{\Gamma})$ [resp. $\tilde{\theta} = (\tilde{\rho}, \tilde{\sigma}, \tilde{\boldsymbol{\Lambda}}, \tilde{\boldsymbol{\Gamma}})$]. We assume that $\frac{\sigma^2}{1-\rho^2} = 1$ and $\boldsymbol{\Lambda}$ is full ranked, with the same constraints holding true for $\tilde{\theta}$. We also assume that $\{\boldsymbol{Y_t}\}$ and $\{\tilde{\boldsymbol{Y_t}}\}$ have the same second-order structure. We prove below that if these conditions hold true then $\theta = \tilde{\theta}$ up to the sign of $\boldsymbol{\Lambda}$ *i.e.* $\rho = \tilde{\rho}$, $\sigma = \tilde{\sigma}$, $\boldsymbol{\Lambda} = \pm\tilde{\boldsymbol{\Lambda}}$ and $\boldsymbol{\Gamma} = \tilde{\boldsymbol{\Gamma}}$. The proof is based on the properties of $\boldsymbol{C}_k = \mathrm{cov}(\boldsymbol{Y_t}, \boldsymbol{Y}_{t+k})$.

- **Identification of ρ and σ.** According to (2.4), we have $\boldsymbol{C}_k = \rho^{k-2}\boldsymbol{C}_2$

for $k \geq 2$ and

$$\boldsymbol{C}_2 = \frac{\sigma^2}{1-\rho^2}\boldsymbol{u}\boldsymbol{v}^t$$

with $\boldsymbol{u} = \boldsymbol{\alpha}_1 + \rho\boldsymbol{\alpha}_0 + \rho^2\boldsymbol{\alpha}_{-1}$ and $\boldsymbol{v} = \rho^2\boldsymbol{\alpha}_1 + \rho\boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_{-1}$. Since $\boldsymbol{\alpha}_{-1}$, $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1$ are linearly independent, $\boldsymbol{u} \neq 0$ and $\boldsymbol{v} \neq 0$ and thus $\boldsymbol{C}_2 \neq 0$. ρ can thus be expressed as a ratio between some coefficients of $\boldsymbol{C}_3$ and $\boldsymbol{C}_2$ and we deduce that $\rho = \tilde{\rho}$. Using the constraint $\frac{\sigma^2}{1-\rho^2} = 1$, we also deduce that $\sigma^2 = \tilde{\sigma}^2$.

- **Identification of $\boldsymbol{\Lambda}$ when $\rho \neq 0$.** According to (2.3-2.4) we have $\boldsymbol{C}_2 - \rho\boldsymbol{C}_1 = (1-\rho^2)\boldsymbol{\alpha}_1\boldsymbol{\alpha}_{-1}^t$ and thus $\boldsymbol{\alpha}_1\boldsymbol{\alpha}_{-1}^t = \tilde{\boldsymbol{\alpha}}_1\tilde{\boldsymbol{\alpha}}_{-1}^t$ since $\rho^2 \neq 1$. We deduce that there exists a real constant $k_1 \neq 0$ such that $\boldsymbol{\alpha}_{-1} = k_1\tilde{\boldsymbol{\alpha}}_{-1}$ and $\boldsymbol{\alpha}_1 = k_1^{-1}\tilde{\boldsymbol{\alpha}}_1$. We also have $\boldsymbol{u}\boldsymbol{v}^t = \tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}^t$ where $\tilde{\boldsymbol{u}}$ and $\tilde{\boldsymbol{v}}$ are defined similarly to $\boldsymbol{u}$ and $\boldsymbol{v}$. We deduce that there exists a real constant $k_2 \neq 0$ such that $\tilde{\boldsymbol{u}} = k_2\boldsymbol{u}$ and $\tilde{\boldsymbol{v}} = k_2^{-1}\boldsymbol{v}$ and thus $\tilde{\boldsymbol{u}} - \tilde{\boldsymbol{v}} = k_2\boldsymbol{u} - k_2^{-1}\boldsymbol{v}$ with

$$\begin{aligned}
\tilde{\boldsymbol{u}} - \tilde{\boldsymbol{v}} &= (1-\rho^2)\tilde{\boldsymbol{\alpha}}_1 + (\rho^2-1)\tilde{\boldsymbol{\alpha}}_{-1} \\
&= (1-\rho^2)k_1^{-1}\boldsymbol{\alpha}_1 + (\rho^2-1)k_1\boldsymbol{\alpha}_{-1}, \text{ and} && (2.6)\\
k_2\boldsymbol{u} - k_2^{-1}\boldsymbol{v} &= (k_2 - \rho^2 k_2^{-1})\boldsymbol{\alpha}_1 + \rho(k_2 - k_2^{-1})\boldsymbol{\alpha}_0 && (2.7)\\
&\quad +(k_2\rho^2 - k_2^{-1})\boldsymbol{\alpha}_{-1} && (2.8)
\end{aligned}$$

Since $\boldsymbol{\alpha}_{-1}$, $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1$ are linearly independent, we can identify the coefficients of the linear combinations (2.6-2.7) and deduce, when $\rho \neq 0$ that $k_2 \in \{-1, 1\}$ and $\boldsymbol{\alpha}_i = k_2\tilde{\boldsymbol{\alpha}}_i$ for $i \in \{-1, 0, 1\}$.

- **Identification of $\boldsymbol{\Lambda}$ when $\rho = 0$.** In this case,

$$\begin{aligned}
\boldsymbol{C_1} &= \sigma^2(\boldsymbol{\alpha}_1\boldsymbol{\alpha}_0^t + \boldsymbol{\alpha}_0\boldsymbol{\alpha}_{-1}^t), && (2.9)\\
\boldsymbol{C_2} &= \sigma^2\boldsymbol{\alpha}_1\boldsymbol{\alpha}_{-1}^t && (2.10)
\end{aligned}$$

By similar reasoning as previously from (2.10) there exists $k_1 \neq 0$ such that $\boldsymbol{\alpha}_{-1} = k_1\tilde{\boldsymbol{\alpha}}_{-1}$ and $\boldsymbol{\alpha}_1 = k_1^{-1}\tilde{\boldsymbol{\alpha}}_1$. From (2.9) we deduce that $\boldsymbol{\alpha}_1(k_1\tilde{\boldsymbol{\alpha}}_0 - \boldsymbol{\alpha}_0)^t + (\frac{\tilde{\boldsymbol{\alpha}}_0}{k_1} - \boldsymbol{\alpha}_0)\boldsymbol{\alpha}_{-1}^t = 0$.

If $k_1\tilde{\boldsymbol{\alpha}}_0 - \boldsymbol{\alpha}_0 \neq 0$ then there exists $k_2 \neq 0$ such that $\boldsymbol{\alpha}_1 - \frac{k_2}{k_1}\tilde{\boldsymbol{\alpha}}_0 - k_2\boldsymbol{\alpha}_0 = 0$ (R_1) and $\frac{1}{k_2}\boldsymbol{\alpha}_{-1} + \boldsymbol{\alpha}_0 + k_1\tilde{\boldsymbol{\alpha}}_0 = 0$ (R_2). Then

$$(\mathrm{R}_1) - \frac{k_2}{k_1}(\mathrm{R}_2) = \boldsymbol{\alpha}_1 + (k_2 + \frac{k_2}{k_1^2})\boldsymbol{\alpha}_0 + \frac{1}{k_1^2}\boldsymbol{\alpha}_{-1} = 0.$$

Since $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_{-1}$ are linearly independent we obtain $k_1 = k_2 = 0$ which is a contradiction.

If $k_1\tilde{\boldsymbol{\alpha}}_0 - \boldsymbol{\alpha}_0 = 0$, this implies $\frac{\tilde{\boldsymbol{\alpha}}_0}{k_1} - \boldsymbol{\alpha}_0 = 0$, then $k_1 = \pm 1$. In both cases, $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_0$ and then identifiable from the covariance $\boldsymbol{C_2}$ and $\boldsymbol{C_1}$.

- **Identification of $\boldsymbol{\Gamma}$.** According to (2.2), $\boldsymbol{\Gamma}$ can be expressed from $\boldsymbol{C_0}$ and the other parameters. We easily deduce that $\tilde{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}$

Here we prove that full-symmetry can not be achieved under the chosen identifiability constraints. Separability of a space-time covariance function implies full-symmetry of this latter (Gneiting, 2002). Full-symmetry of the space-time covariance function implies that the matrix $\boldsymbol{C_2}$ is a symmetric matrix. The symmetry of $\boldsymbol{C_2}$ implies $\boldsymbol{u}\boldsymbol{v}^t = \boldsymbol{v}\boldsymbol{u}^t$, $\boldsymbol{u}$ and $\boldsymbol{v}$ are then collinear vectors which implies a collinearity between $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_{-1}$. The space-time covariance function defined by the model is not fully-symmetric and then non-separable.

2.8 Maximum Likelihood Estimation for the model (M) and associated reduced models

Maximum likelihood estimation of the parameter θ for models with latent variables consists in maximizing the incomplete likelihood function based on observed set $(\boldsymbol{y}_1, ..., \boldsymbol{y}_T)$:

$$\begin{aligned}\mathcal{L}(\theta; \boldsymbol{y}_1, ..., \boldsymbol{y}_T) &= p(\boldsymbol{y}_1, ..., \boldsymbol{y}_T; \theta) \\ &= \mathcal{L}(\theta; \boldsymbol{y}_1, ..., \boldsymbol{y}_T) = p(\boldsymbol{y}_1) \prod_{t=2}^{T} p(\boldsymbol{y}_t | \boldsymbol{y}_1, ..., \boldsymbol{y}_{t-1}; \theta).\end{aligned}$$

In the Gaussian linear case, the likelihood of the observations $(\boldsymbol{y}_1, ..., \boldsymbol{y}_T)$ can be computed easily since for all $t \in \{1, ..., T\}$ $(\boldsymbol{Y}_1, ..., \boldsymbol{Y}_t)$ is a Gaussian vector. It gives for the model (M):

$$\mathcal{L}(\boldsymbol{Y}_t | \boldsymbol{Y}_1^{t-1} = \boldsymbol{y}_1^{t-1}) = \mathcal{N}(\boldsymbol{\Lambda}\tilde{\boldsymbol{X}}_{t|t-1}, \boldsymbol{F}_{t|t-1}) \text{ where } \tilde{\boldsymbol{X}}_t = \begin{pmatrix} X_{t+1} \\ X_t \\ X_{t-1} \end{pmatrix},$$

with $\tilde{\boldsymbol{X}}_{t|t-1} = \mathrm{E}(\tilde{\boldsymbol{X}}_t | \boldsymbol{Y}_1^{t-1} = \boldsymbol{y}_1^{t-1})$ and $\boldsymbol{F}_{t|t-1} = \mathrm{var}(\boldsymbol{Y}_t | \boldsymbol{Y}_1^{t-1} = \boldsymbol{y}_1^{t-1}) = \boldsymbol{\Lambda}\boldsymbol{P}_{t|t-1}\boldsymbol{\Lambda}^t + \boldsymbol{\Gamma}$, where $\boldsymbol{P}_{t|t-1} = \mathrm{var}(\tilde{\boldsymbol{X}}_t | \boldsymbol{Y}_1^{t-1} = \boldsymbol{y}_1^{t-1}) = \mathrm{E}((\tilde{\boldsymbol{X}}_t - \tilde{\boldsymbol{X}}_{t|t-1})(\tilde{\boldsymbol{X}}_t - \tilde{\boldsymbol{X}}_{t|t-1})^t | \boldsymbol{Y}_1^{t-1} = \boldsymbol{y}_1^{t-1})$ with $\boldsymbol{y}_1^{t-1} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_{t-1})$. Both quantities $\tilde{\boldsymbol{X}}_{t|t-1}$ and $\boldsymbol{P}_{t|t-1}$ are computed from Kalman filter described below (see also (Shumway and Stoffer, 2006)). However no explicit expressions of the optimal parameters are available from this incomplete likelihood, a maximum likelihood estimation procedure would involve a numerical optimization of this function which is not reasonable in high dimension. A major feature of the EM algorithm (Dempster et al., 1977) is the maximization of the complete likelihood over the parameter θ.

2.8.1 Kalman recursions

The goal of filtering (respectively smoothing, respectively prediction) is to obtain as much as possible information about the hidden variable X_t from the observations $(\boldsymbol{y}_1, ..., \boldsymbol{y}_t)$ (respectively $(\boldsymbol{y}_1, ..., \boldsymbol{y}_T)$, respectively $(\boldsymbol{y}_1, ..., \boldsymbol{y}_{t-1})$). The solution consists in computing recursively the conditional law of X_t according to $(\boldsymbol{y}_1, ..., \boldsymbol{y}_t)$ (respectively $(\boldsymbol{y}_1, ..., \boldsymbol{y}_T)$, respectively $(\boldsymbol{y}_1, ..., \boldsymbol{y}_{t-1})$), which realizes the best approximation of X_t according to $(\boldsymbol{y}_1, ..., \boldsymbol{y}_t)$ in terms of mean square error.

Kalman prediction and filtering: $(\tilde{\boldsymbol{X}}_t, \boldsymbol{Y}_1, ..., \boldsymbol{Y}_{t-1})$ is a Gaussian vector then the conditional distribution of $\tilde{\boldsymbol{X}}_t$ according to $(\boldsymbol{Y}_1 = \boldsymbol{y}_1, ..., \boldsymbol{Y}_{t-1} = \boldsymbol{y}_{t-1})$ is a Gaussian distribution with parameters: $\tilde{\boldsymbol{X}}_{t|t-1} = E(\tilde{\boldsymbol{X}}_t | \boldsymbol{Y}_1 = \boldsymbol{y}_1, ..., \boldsymbol{Y}_{t-1} = \boldsymbol{y}_{t-1})$ and $\boldsymbol{P}_{t|t-1} = \mathrm{var}(\tilde{\boldsymbol{X}}_t | \boldsymbol{Y}_1 = \boldsymbol{y}_1, ..., \boldsymbol{Y}_{t-1} = \boldsymbol{y}_{t-1}) = \mathrm{E}((\tilde{\boldsymbol{X}}_t - \tilde{\boldsymbol{X}}_{t|t-1})(\tilde{\boldsymbol{X}}_t - \tilde{\boldsymbol{X}}_{t|t-1})^t | \boldsymbol{Y}_1^{t-1} = \boldsymbol{y}_1^{t-1})$; and $\mathcal{L}(\tilde{\boldsymbol{X}}_t | \boldsymbol{Y}_1 = \boldsymbol{y}_1, ..., \boldsymbol{Y}_t = \boldsymbol{y}_t) = \mathcal{N}(\boldsymbol{\Lambda}\tilde{\boldsymbol{X}}_{t|t}, \boldsymbol{P}_{t|t})$. Relationships between predicted and filtered quantities are the following:

$$\tilde{\boldsymbol{X}}_{t|t-1} = \tilde{\boldsymbol{\rho}}\tilde{\boldsymbol{X}}_{t-1|t-1},$$

$$\tilde{\boldsymbol{P}}_{t|t-1} = \tilde{\boldsymbol{\rho}}\tilde{\boldsymbol{P}}_{t-1|t-1}\tilde{\boldsymbol{\rho}}^t + \tilde{\boldsymbol{\sigma}},$$

$$\tilde{\boldsymbol{X}}_{t|t} = \tilde{\boldsymbol{X}}_{t|t-1} + \boldsymbol{K}_t(\boldsymbol{Y}_t - \boldsymbol{\Lambda}\tilde{\boldsymbol{X}}_{t|t-1})$$

and

$$\tilde{\boldsymbol{P}}_{t|t} = (I - \boldsymbol{K}_t\boldsymbol{\Lambda})\tilde{\boldsymbol{P}}_{t|t-1},$$

where $\boldsymbol{K}_t = \tilde{\boldsymbol{P}}_{t|t-1}\boldsymbol{\Lambda}^t(\boldsymbol{\Lambda}\tilde{\boldsymbol{P}}_{t|t-1}\boldsymbol{\Lambda}^t + \boldsymbol{\Gamma})^{-1}$ and $\boldsymbol{K}$ is called the Kalman gain. The two first expressions are easily derived from independence of $\boldsymbol{\epsilon}_t$ and $\boldsymbol{Y}_{t-1}$ and of $(\tilde{\boldsymbol{X}}_{t-1} - \tilde{\boldsymbol{X}}_{t-1|t-1})$ and $\boldsymbol{\epsilon}_t$. The two last relations are based on properties of the Gaussian process of innovations $\boldsymbol{I}_t = \boldsymbol{Y}_t - \mathrm{E}(\boldsymbol{Y}_t | \boldsymbol{Y}_1 = \boldsymbol{y}_1, ..., \boldsymbol{Y}_{t-1} = \boldsymbol{y}_{t-1})$.

Kalman smoothing: Computation of $\tilde{\boldsymbol{X}}_{t|t}$ and $\tilde{\boldsymbol{P}}_{t|t}$ is obtained through the following backward recursions:

$$\tilde{\boldsymbol{X}}_{t|t} = \tilde{\boldsymbol{X}}_{t|t} + \boldsymbol{J}_t(\tilde{\boldsymbol{X}}_{t+1|T} - \tilde{\boldsymbol{\rho}}\tilde{\boldsymbol{X}}_{t|t}),$$

$$\tilde{\boldsymbol{P}}_{t|t} = \tilde{\boldsymbol{P}}_{t|t} + \boldsymbol{J}_t(\tilde{\boldsymbol{P}}_{t+1|T} - \tilde{\boldsymbol{P}}_{t+1|t})\boldsymbol{J}_t^t,$$

$$\boldsymbol{J}_t = \tilde{\boldsymbol{P}}_{t|t}\tilde{\boldsymbol{\rho}}^t\tilde{\boldsymbol{P}}_{t+1|t}^{-1}$$

and

$$\tilde{\boldsymbol{P}}_{t,t-1|T} = \tilde{\boldsymbol{P}}_{t|t}J_{t-1}^t + \boldsymbol{J}_t(\tilde{\boldsymbol{P}}_{t+1,t|T} - \tilde{\boldsymbol{\rho}}\tilde{\boldsymbol{P}}_{t|t})J_{t-1}^t.$$

Similar computations of the previous ones based on conditional expectation of multivariate normal distribution are used to compute these quantities.

2.8.2 EM algorithm

Thanks to the Markov properties and Bayes formula, the complete likelihood of the model (M) for $\theta = (\rho, \sigma, \mathbf{\Lambda}, \mathbf{\Gamma})$ is written as:

$$\begin{aligned}\mathcal{L}(\theta; x_0, ..., x_T, \boldsymbol{y}_1, ..., \boldsymbol{y}_T) &= \mathcal{L}(\theta; \tilde{\boldsymbol{X}}_1, ..., \tilde{\boldsymbol{X}}_{T-1}, \boldsymbol{y}_1, ..., \boldsymbol{y}_T) \\ &= p(x_0) \prod_{i=1}^{T} p(x_i | x_{i-1}; \theta) \prod_{i=1}^{T} p(\boldsymbol{y}_i | \tilde{\boldsymbol{X}}_i; \theta).\end{aligned}$$

However the set $(x_0, ..., x_T)$ is not observed, the EM-algorithm enables to approximate $\hat{\theta}$ that maximizes the quantity $\mathrm{E}(\log(p(\tilde{\boldsymbol{X}}_1, ..., \tilde{\boldsymbol{X}}_t, \boldsymbol{Y}_1, ..., \boldsymbol{Y}_T; \theta)) | \boldsymbol{Y}_1^T = \boldsymbol{y}_1^T)$. The EM-algorithm computes approximations $\hat{\theta}_n$ of $\hat{\theta}$ in a recursive way by performing the following two steps at each iteration n:

Expectation step: Computation of

$$Q(\theta, \hat{\theta}_n) = \mathrm{E}(\log(\mathcal{L}(\tilde{\boldsymbol{X}}_1, ..., \tilde{\boldsymbol{X}}_{T-1}, \boldsymbol{Y}_1, ..., \boldsymbol{Y}_T; \theta)) | \boldsymbol{Y}_1^T = \boldsymbol{y}_1^T; \hat{\theta}_n),$$

through the Kalman filtering and smoothing recursions (see (Shumway and Stoffer, 2006)).

Maximization step: Computation of $\hat{\theta}_{n+1}$ by maximization of the function $(\theta \to Q(\theta, \hat{\theta}_n))$.

Since $\boldsymbol{X}^t \boldsymbol{M} \boldsymbol{X} = \mathrm{Trace}(\boldsymbol{M} \boldsymbol{X} \boldsymbol{X}^t)$ for all K-dimensional vector $\boldsymbol{X}$ and $K \times K$-matrix $\boldsymbol{M}$, the quantity $Q(\theta, \hat{\theta}_n)$ is derived:

$$\begin{aligned}Q(\theta, \hat{\theta}_n) = &-\frac{1}{2}\Big((T-1)(\log(2\pi) + \log(\sigma^2)) \\ &+ \frac{1}{\sigma^2} \sum_{i=2}^{T} \mathrm{E}((X_i - \rho X_{i-1})^2 | \boldsymbol{y}_1^T; \hat{\theta}_n) + T(K \log(2\pi) + \log(\det(\mathbf{\Gamma}))) \\ &+ \sum_{i=1}^{T} \mathrm{Trace}(\mathbf{\Gamma}^{-1} \mathrm{E}((\boldsymbol{y}_i - \mathbf{\Lambda} \tilde{\boldsymbol{X}}_i)(\boldsymbol{y}_i - \mathbf{\Lambda} \tilde{\boldsymbol{X}}_i)^t | \boldsymbol{y}_1^T; \hat{\theta}_n))\Big).\end{aligned}$$

Then the following quantities $\hat{x}_i = \mathrm{E}(X_i | \boldsymbol{y}_1^T; \hat{\theta}_n)$, $\hat{x}_{i,i-1} = \mathrm{E}(X_i X_{i-1}{}^t | \boldsymbol{y}_1^T; \hat{\theta}_n)$, $\hat{\tilde{\boldsymbol{X}}}_i = \mathrm{E}(\tilde{\boldsymbol{X}}_i | \boldsymbol{y}_1^T; \hat{\theta}_n)$ and $\hat{\tilde{\boldsymbol{X}}}_{i,i} = \mathrm{E}(\tilde{\boldsymbol{X}}_i \tilde{\boldsymbol{X}}_i{}^t | \boldsymbol{y}_1^T; \hat{\theta}_n)$ are needed for all $i \in \{1, ..., T\}$ and derived from the Kalman filter and smoother. At each M-step, analytical expressions of the estimates of the parameters can be derived:

$$\rho_n = \frac{\sum_{i=2}^{T} \hat{x}_{i,i-1}}{\sum_{i=1}^{T} \hat{x}_{i,i}},$$

$$\mathbf{\Lambda}_n = \Big(\sum_{i=1}^{T} \boldsymbol{y}_i \hat{\tilde{\boldsymbol{X}}}_i{}^t\Big)\Big(\sum_{i=1}^{T} \hat{\tilde{\boldsymbol{X}}}_{i,i}\Big)^{-1} \text{ and } \mathbf{\Gamma}_n = \frac{1}{T}\sum_{i=1}^{T}(\boldsymbol{y}_i\boldsymbol{y}_i{}^t - \mathbf{\Lambda}_n \hat{\tilde{\boldsymbol{X}}}_i{}^t \boldsymbol{y}_i{}^t).$$

The estimation of $\mathbf{\Gamma}_n$ in models (M_Γ) and of $\mathbf{\Lambda}_n$ in the model (M_Λ) are processed by numerical optimization of the associated part of the log-likelihood. For the model (M_Γ), $\mathbf{\Lambda}_n$ is determined by its analytical expression and injected in the associated part of the likelihood which is optimized numerically to determine the parameters that structure $\mathbf{\Gamma}_n$. $\mathbf{\Gamma}_n$ is the maximizer of:

$$\begin{aligned}(\sigma_1, ..., \sigma_K, \mathbf{\Lambda}_1, \mathbf{\Lambda}_2) \rightarrow \quad & T(K\log(2\pi) + \log(\det(\mathbf{\Gamma}_{par}))) + \\ & \sum_{i=1}^{T} \mathrm{Trace}(\mathbf{\Gamma}_{par}^{-1}\mathrm{E}((\boldsymbol{y}_i - \mathbf{\Lambda}_n\tilde{\boldsymbol{X}}_i)(\boldsymbol{y}_i - \mathbf{\Lambda}_n\tilde{\boldsymbol{X}}_i)^t|\boldsymbol{y}_1^T;\hat{\theta}_n)).\end{aligned}$$

Where $\mathbf{\Gamma}_{par}$ is the parametric covariance defined by $(\sigma_1, ..., \sigma_K, \mathbf{\Lambda}_1, \mathbf{\Lambda}_2)$. Initial conditions of the parameters of the structure of $\mathbf{\Gamma}$ are determined empirically. In the estimation procedure associated with (M_Λ), $\mathbf{\Lambda}_n$ is determined as the maximizer of the function:

$$\begin{aligned}(\beta_1^{\mathrm{Lat}}, ..., \beta_9) \rightarrow \quad & T(K\log(2\pi) + \log(\det(\mathbf{\Gamma}_{n-1}))) + \\ & \sum_{i=1}^{T} \mathrm{Trace}(\mathbf{\Gamma}_{n-1}^{-1}\mathrm{E}((\boldsymbol{y}_i - \mathbf{\Lambda}_{par}\tilde{\boldsymbol{X}}_i)(\boldsymbol{y}_i - \mathbf{\Lambda}_{par}\tilde{\boldsymbol{X}}_i)^t|\boldsymbol{y}_1^T;\hat{\theta}_n)),\end{aligned}$$

with $\mathbf{\Lambda}_{par} = \begin{pmatrix} 1 & | & \mathrm{Long} & | & \mathrm{Long}^2 \end{pmatrix} \begin{pmatrix} \beta_1^{\mathrm{Lat}} & \beta_4^{\mathrm{Lat}} & \beta_7^{\mathrm{Lat}} \\ \beta_2 & \beta_5 & \beta_8 \\ \beta_3 & \beta_6 & \beta_9 \end{pmatrix}$. Initial conditions of this optimization are determined by a least square estimation between $\hat{\mathbf{\Lambda}}$, the output of the EM processes for the model (M), and $\mathbf{\Lambda}_{par}$. $\mathbf{\Gamma}_n$ is then determined as the maximizer of:

$$\begin{aligned}\mathbf{\Gamma} \rightarrow \quad & T(K\log(2\pi) + \log(\det(\mathbf{\Gamma}))) \\ & + \sum_{i=1}^{T} \mathrm{Trace}(\mathbf{\Gamma}^{-1}\mathrm{E}((\boldsymbol{y}_i - \mathbf{\Lambda}_n\tilde{\boldsymbol{X}}_i)(\boldsymbol{y}_i - \mathbf{\Lambda}_n\tilde{\boldsymbol{X}}_i)^t|\boldsymbol{y}_1^T;\hat{\theta}_n)).\end{aligned}$$

The splitting of optimization in $\mathbf{\Lambda}$ and $\mathbf{\Gamma}$ into the EM algorithm refers to a Generalized Expectation-Maximization algorithm in which at each M-step only an improvement of the approximated incomplete likelihood is required.

2.9 Prediction as a validation tool

The time-step of the data makes unrealistic the use of the proposed model as a forecasting tool. Nevertheless, forecasting is used here a classical statistical tool for validation. Indeed it enables to evaluate many features linked to

statistical modeling and it can, for instance, help to detect overfitting. The Markovian structure of the model (M) is such that the short-term forecast can be efficiently computed through the Kalman recursions (see (Brockwell and Davis, 2006, chapter 8)). The forecast is performed on the last 8 years of data (validation set) after fitting the model on the first 25 years of data (training set). In practice the forecast skills of the model at location $i \in \{1, ..., K\}$ is evaluated by computing the natural empirical estimate of the Mean Square Percentage Error (MSPE) defined as

$$\mathrm{MSPE}(i) = \frac{\mathrm{var}(\boldsymbol{Y_t}(i) - \mathrm{E}[\boldsymbol{Y_t}(i)|\boldsymbol{Y}_0, ..., \boldsymbol{Y}_{t-1}])}{\mathrm{var}(\boldsymbol{Y_t}(i))}$$

where the MSE of the forecast error (the numerator) is normalized by the variance of the field at the individual locations, with $\boldsymbol{Y_t}$ the original non transformed wind.

For comparison purpose, a vector autoregressive model of order 1 (VAR(1)) was also fitted on the multivariate process $\boldsymbol{Y}$ of transformed mean-corrected wind speed. Such a high-dimension response vector may lead to a model VAR which suffers from over-parameterization and to a difficult interpretation of the parameters. Note that the BIC and MSPE criteria lead to coherent results.

Chapter 3

Non-homogeneous hidden Markov-Switching AutoRegressive models for wind time series

This chapter is the object of a submitted paper:
Ailliot, P.,Bessac, J., Monbet, V. and Pène, F. (2014). Non-homogeneous hidden Markov-Switching AutoRegressive models for wind time series. Submitted

In this work, 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 permit to reproduce complex features of wind time series such as non-linear dynamics and the multi-modal marginal distributions.

3.1 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 (Srikanthan and McMahon, 1999) and references therein). These generators can also be useful for downscaling global climate models (see e.g. (Maraun et al., 2010) and references therein) or infilling missing values by conditional simulation. In this work we focus on wind time series. Various approaches have been proposed in the literature for modeling time series of wind speed (see e.g. (Monbet et al., 2007) 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 work 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 3.1). 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 (Vautard, 1990) 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 (Zucchini and Guttorp, 1991) 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 (Zucchini and MacDonald, 2009). However HMMs assume that the successive observations are conditionally independent given the latent weather type and cannot reproduce 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) models have been proposed in this context to model time series of wind speed in (Monbet et al., 2007; Pinson et al., 2008; Ailliot and Monbet, 2012). 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. (Brown et al., 1984)), 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 work 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 work and a discussion of their respective advantages is given.

The work is organized as follows. NHMS-AR models are introduced in Section 4.2 with specific parameterizations proposed when considering Cartesian and polar coordinates. In Section 3.3, 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 3.4. The data used in this work 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 3.5.

3.2 Models

3.2.1 Non-homogeneous Markov-switching autoregressive models

Let $X_t \in \{1, ..., M\}$ represent the latent weather type and Y_t denote 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 3.2.3, the wind direction in Section 3.2.4, and finally $\{Y_t\}$ stands for the wind speed in Section 3.2.5. 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 $Y_t^{t+u} = (Y_t, ..., Y_{t+u})$, $y_t^{t+u} = (y_t, ..., y_{t+u})$ (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 $\{1, ..., M\} \times \mathbb{E}$ such that*

- *the conditional distribution of X_t given the values of $\{X_{t'}\}_{t'<t}$ and $\{Y_{t'}\}_{t'<t}$*

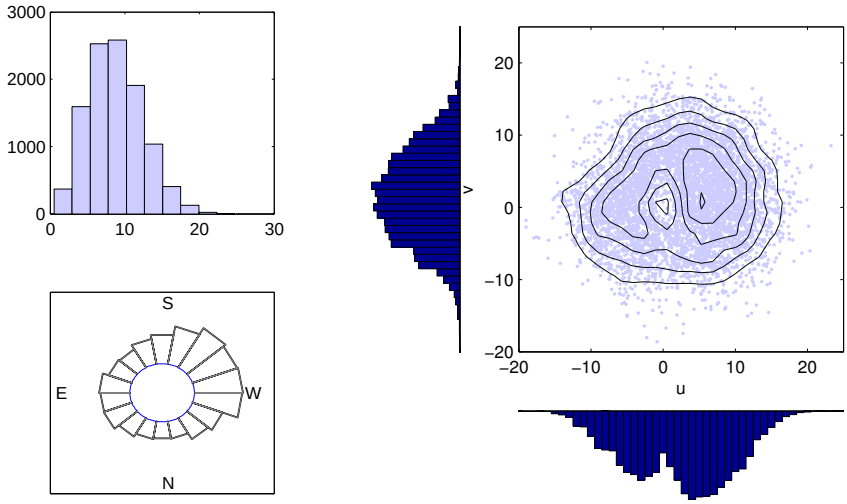

Figure 3.1: Histogram of $\{U_t\}$ (top left), rose plot of $\{\Phi_t\}$ (bottom left) and histograms of $\{u_t\}$ and $\{v_t\}$ and joint distribution of $\{u_t, v_t\}$ (right). The lines on the scatter plots are levels of a non-parametric kernel estimate of the bivariate density. Results for the months of January.

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 $Y_{t-1}, \ldots, Y_{t-s}$ and this conditional distribution has a probability density function (p.d.f.) $p_2\left(y_t|x_t, y_{t-s}^{t-1}\right)$.*

Let us write $p(.|x_{t-u}^{t-1}, y_{t-u}^{t-1})$ 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 $u \geq s$

$$p(x_t, y_t|x_{t-u}^{t-1}, y_{t-u}^{t-1}) = p_1(x_t|x_{t-1}, y_{t-1})p_2(y_t|x, y_{t-s}^{t-1}). \tag{3.1}$$

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

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

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. (Tong, 1990)).

The following sections propose specific parametric models for p_1 (see Section 3.2.2) and p_2 when using Cartesian coordinates (see 3.2.3) or polar coordinates (see Sections 3.2.4 and 3.2.5).

3.2.2 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 that when it is blowing from the South. Such features can be modeled through the transition kernel p_1. Hereafter we assume that

$$p_1(x_t|x_{t-1}, \phi_{t-1}) \propto q_{x_{t-1},x_t} f_{VM}(\phi_{t-1}; \lambda_{x_{t-1},x_t}, \psi_{x_{t-1},x_t}), \tag{3.2}$$

where $f_{VM}(.; \kappa, \phi)$ is the probability density function (p.d.f) of the von Mises distribution, ϕ_{t-1} is the wind direction at time $t-1$, $Q = (q_{x,x'})_{x,x' \in \{1,...,M\}}$ is a stochastic matrix and, for $x, x' \in \{1, ..., M\}$, $\lambda_{x,x'} \geq 0$ and $\psi_{x,x'} \in \mathbb{T}$ are unknown parameters. The von Mises distribution is a natural distribution for circular variables (see (Mardia, 1972)) which p.d.f. with respect to the Lebesgue measure on $\mathbb{T}$, is given by

$$\forall z \in \mathbb{T}, \;\; f_\gamma(z) = f_{VM}(z; \kappa, \phi) = \frac{1}{2\pi I_0(\kappa)} \exp\big(\kappa \cos(z - \phi)\big) = \frac{1}{2\pi I_0(|\gamma|)} \left| e^{\gamma e^{-iz}} \right|, \tag{3.3}$$

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 (3.2), 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 $x, x' \in \{1, ..., M\}$ then we obtain again the homogeneous MS-AR models. Observe that (3.2) can be rewritten

$$p_1(x_t|x_{t-1}, \phi_{t-1}) = \frac{q_{x_{t-1},x_t} \left| \exp\left(\tilde{\lambda}_{x_{t-1},x_t} e^{-i\phi_{t-1}}\right) \right|}{\sum_{x'=1}^{M} q_{x_{t-1},x'} \left| \exp\left(\tilde{\lambda}_{x_{t-1},x'} e^{-i\phi_{t-1}}\right) \right|}, \tag{3.4}$$

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

$$\tilde{\lambda}_{x,x'} = \tilde{\lambda}_{x'} \tag{3.5}$$

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

$$\sum_{x'=1}^{M} \tilde{\lambda}_{x'} = 0. \tag{3.6}$$

We have also fitted the model without the constraint (3.5) and found that the likelihood of these models is similar to the one of the models with the constraint (3.5) whereas they have a significantly larger number of parameters. Even when assuming (3.5), 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.

3.2.3 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

$$Y_t = A_0^{(x_t)} + A_1^{(x_t)} Y_{t-1} + ... + A_s^{(x_t)} Y_{t-s} + \left(\Sigma^{(x_t)}\right)^{\frac{1}{2}} \epsilon_t \tag{3.7}$$

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

VAR models have been proposed for wind fields in a space-time context in (Ailliot et al., 2006b; Wikle et al., 2001; Fuentes et al., 2005). 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 3.1. 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 (3.2), is denoted **NHMS-AR**$_{(u,v)}$.

3.2.4 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 (Breckling, 1989; Fisher and Lee, 1994; Holzmann et al., 2006; Kato, 2010)). We have chosen to focus on the von Mises process initially introduced in (Breckling, 1989) and assume that the conditional distribution of Y_t given $\left(X_t = x_t, Y_{t-s}^{t-1} = y_{t-s}^{t-1}\right)$ 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 $x \in \{1, ..., M\}$ and $\ell \in \{0, ..., s\}$. This can be rewritten

$$\begin{aligned} & p_2(y_t | x_t, y_{t-s}^{t-1}) \\ = & \frac{1}{b(x_t, y_{t-s}^{t-1})} \exp\left(\kappa_0^{(x_t)} \cos(y_t - \phi_0^{(x_t)}) + \sum_{\ell=1}^{s} \kappa_\ell^{(x_t)} \cos(y_t - y_{t-\ell} - \phi_\ell^{(x)})\right) \\ = & \frac{1}{b(x_t, y_{t-s}^{t-1})} \left| \exp\left([\gamma_0^{(x_t)} + \sum_{\ell=1}^{s} \gamma_\ell^{(x_t)} e^{iy_{t-\ell}}] e^{-iy_t}\right) \right| \end{aligned} \tag{3.8}$$

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 (Breckling, 1989), it was assumed that $\gamma_\ell^{(x)} \in \mathbb{R}$ for $\ell \in \{1, ..., s\}$. 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 3.4).

In the sequel, the model with homogeneous hidden Markov chain is denoted **HMS-EVM** and the non-homogeneous model, with p_1 is given by (3.2), is written **NHMS-EVM**.

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

In (Ailliot and Monbet, 2012) 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 (3.7)) of order $s = 2$. Figure 3.2 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 3.3 (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 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

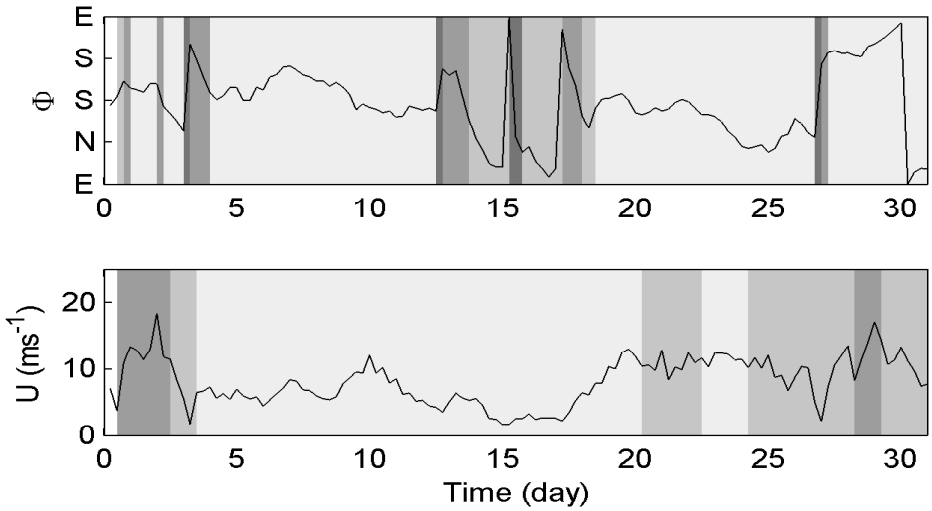

Figure 3.2: Example of time series of wind direction (top plot) and wind speed (bottom plot). The colors indicate the most likely regimes for the fitted **NHMS-EVM** model with 4 regimes (top plot) and Gaussian homogeneous MS-AR model for wind speed with 3 regimes (bottom plot). The regimes have been ordered according to the time variability (the darker the more variability).

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 (3.2) and p_2 by a linear Gaussian AR model (3.7).

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.

$$
\begin{tikzcd}
\cdots \arrow[r] \arrow[d] & X_{t-1}^{(\Phi)} \arrow[r] \arrow[d] & X_t^{(\Phi)} \arrow[r] \arrow[d] & X_{t+1}^{(\Phi)} \arrow[r] \arrow[d] & \cdots \arrow[d] \\
\cdots \arrow[r] \arrow[ur] \arrow[dr] & \Phi_{t-1} \arrow[r] \arrow[ur] \arrow[dr] & \Phi_t \arrow[r] \arrow[ur] \arrow[dr] & \Phi_{t+1} \arrow[r] \arrow[ur] \arrow[dr] & \cdots \\
\cdots \arrow[r] \arrow[d] & X_{t-1}^{(U)} \arrow[r] \arrow[d] & X_t^{(U)} \arrow[r] \arrow[d] & X_{t+1}^{(U)} \arrow[r] \arrow[d] & \cdots \arrow[d] \\
\cdots \arrow[r] & U_{t-1} \arrow[r] & U_t \arrow[r] & U_{t+1} \arrow[r] & \cdots
\end{tikzcd}
$$

3.3 Parameter estimation

3.3.1 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 $x \in \{1, ..., M\}$.

They are estimated by maximizing the likelihood function using a generalized EM algorithm. This algorithm was initially introduced in (Baum et al., 1970) for HMMs and then generalized to models with latent variables in (Dempster et al., 1977). 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 $Q\left(\theta|\hat{\theta}_i\right) = E_{\hat{\theta}_i}(\log(p_\theta(X_1^T, Y_1^T))|y_{-s+1}^T)$ as a function of θ.

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

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. (Cappé et al., 2005) and references therein). The particular implementation of these recursions for homogeneous MS-AR models is discussed in (Hamilton, 1990) and (Hughes et al., 1999) discusses it for non-homogeneous HMMs. It can be easily generalized to the models considered in this work. 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 $Q(.|\hat{\theta}_i)$ which has to be maximized in the M-step can be written as the sum of $M+1$ functions as follows

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

This leads to solving $M+1$ separate optimization problems on spaces with reduced dimension which is far more efficient than maximizing directly $Q(.|\hat{\theta}_i)$ over all the parameters. Note that analytical expressions are available for $\arg\max_{\theta^{(x)}} Q_Y\left(\theta^{(x)}|\hat{\theta}_i\right)$ when Gaussian linear AR models (3.7) are used and for $\arg\max_{\theta_Q} Q_X\left(\theta_Q|\hat{\theta}_i\right)$ 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 $s+1$.

3.3.2 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 (Ailliot and Pène, 2013) for the models with linear Gaussian AR models (3.7). 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 (Ailliot and Pène, 2013)). These conditions apply to the models with linear Gaussian AR discussed in Section 3.4 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 3.2.4 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 (3.4) and (3.8) respectively. Let Θ' be the set $\theta = (q_{x,x'}, \tilde{\lambda}_{x,x'}, \gamma_\ell^{(x)})_{\ell,x,x'}$ 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 (3.6). 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*

$$x \neq x' \;\Rightarrow\; (\gamma_{0,*}^{(x)}, ..., \gamma_{s,*}^{(x)}) \neq (\gamma_{0,*}^{(x')}, ..., \gamma_{s,*}^{(x')}). \tag{3.9}$$

Then, for every $x_0 \in \{1, ..., M\}$ and any initial measure ν on $\{1, ..., M\} \times \mathbb{T}$, 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 $\{1, ..., M\}$ such that, for every $x, x' \in \{1, ..., M\}$, for every $j = 0, ..., s$, the following relations hold true

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

One can notice that (3.9) 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 (Ailliot and Pène, 2013)[Thm 2] and the proof of the identifiability up to a permutation of indices (see Proposition 2). Let us first check that the conditions of (Ailliot and Pène, 2013)[Thm 2] apply for the NHMS-EVM model.

Since for every $(\theta, x, y) \in \Theta \times \{1, ..., M\} \times \mathbb{T}$, $q_\theta(x, y|\cdot, \cdot)$ is continuous on the compact set $\{1, ..., M\} \times \mathbb{T}^s$, 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 (Ailliot and Pène, 2013)[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 with respect to (θ, x_1, x_0, y_0) and wit respect to $(\theta, x_0, y_0, y_{-1})$ (respectively), all the other assumptions of (Ailliot and Pène, 2013)[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 2 (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 (3.8) for θ_1 are such that

$$x \neq x' \;\Rightarrow\; (\gamma_{0,(1)}^{(x)}, ..., \gamma_{s,(1)}^{(x)}) \neq (\gamma_{0,(1)}^{(x')}, ..., \gamma_{s,(1)}^{(x')}). \tag{3.10}$$

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 4 (Proof of Proposition 2) *We write* Leb *for the Lebesgue measure on* $\mathbb{T}$. *Assume that* $\bar{\mathbb{P}}^Y_{\theta_1} = \bar{\mathbb{P}}^Y_{\theta_2}$. *In particular, we have*

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

and thus

$$\begin{aligned}\sum_{x=1}^{M} \bar{\mathbb{P}}_{\theta_1}(X_t = x | y^{t-1}_{t-s}) p_{2,\theta_1}(y_t | x, y^{t-1}_{t-s}) &= \sum_{x=1}^{M} \bar{\mathbb{P}}_{\theta_2}(X_t = x | y^{t-1}_{t-s}) p_{2,\theta_2}(y_t | x, y^{t-1}_{t-s}), \\ &\quad \text{for } \bar{\mathbb{P}}^{Y^t_{t-s}}_{\theta_1} - a.e.\ y^t_{t-s}.\end{aligned}$$

Since $\bar{p}_{\theta_1}(y^t_{t-s}) > 0$ *(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 (3.4) holds true, we deduce that, for* $\mathrm{Leb}^{\otimes(s+1)}$*-a.e.* y^t_{t-s}, *we have*

$$\begin{aligned}&\sum_{x=1}^{M} \bar{\mathbb{P}}_{\theta_1}(X_t = x | y^{t-1}_{t-s}) f_{\gamma^{(x)}_{0,(1)} + \sum_{\ell=1}^{s} \gamma^{(x)}_{\ell,(1)} e^{i y_{t-\ell}}}(y_t) \\ &= \sum_{x=1}^{M} \bar{\mathbb{P}}_{\theta_2}(X_t = x | y^{t-1}_{t-s}) f_{\gamma^{(x)}_{0,(2)} + \sum_{\ell=1}^{s} \gamma^{(x)}_{\ell,(2)} e^{i y_{t-\ell}}}(y_t)\end{aligned}$$

with f_γ *defined by (3.3). Due to (Fraser et al., 1981), finite mixtures of von Mises distributions are identifiable. Hence if*

$$\sum_{x=1}^{M} \pi^{(x)}_1 f_{\gamma^{(x)}_1}(y) = \sum_{x=1}^{M} \pi^{(x)}_2 f_{\gamma^{(x)}_2}(y) \text{ for } \mathrm{Leb} - a.e.\ y$$

with $\gamma^{(x)}_1 \neq \gamma^{(x')}_1$ *for* $x \neq x'$ *and* $\pi^{(x)}_1 > 0$ *for* $x \in \{1, ..., M\}$ *then there exists a permutation* $\tau : \{1, ..., M\} \to \{1, ..., M\}$ *such that* $\gamma^{(x)}_1 = \gamma^{(\tau(x))}_2$ *and* $\pi^{(x)}_1 = \pi^{(\tau(x))}_2$.

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

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

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

$$\forall x \in \{1, ..., M\}, \quad \gamma^{(x)}_{0,(1)} + \sum_{\ell=1}^{s} \gamma^{(x)}_{\ell,(1)} e^{i y_{t-\ell}} = \gamma^{(\sigma_{y^{t-1}_{t-s}}(x))}_{0,(2)} + \sum_{\ell=1}^{s} \gamma^{(\sigma_{y^{t-1}_{t-s}}(x))}_{\ell,(2)} e^{i y_{t-\ell}}.$$

Since the set of permutations of $\{1, ..., M\}$ *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*

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

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

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

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*

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

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

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

Using again the identifiability of von Mises distribution, we obtain

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

Now, due to the special form of $p_{1,\theta}$ *specified in* (3.4)*, we get*

$$\begin{aligned}\forall x, x',\quad \mathrm{Leb}-a.e.\ y_t,\qquad &\frac{q_{x,x',(1)} \left|\exp\left(\tilde{\lambda}_{x,x',(1)} e^{-iy_t}\right)\right|}{\sum_{x''=1}^{M} q_{x,x'',(1)} \left|\exp\left(\tilde{\lambda}_{x,x'',(1)} e^{-iy_t}\right)\right|}\\ &= \frac{q_{\sigma(x),\sigma(x'),(2)} \left|\exp\left(\tilde{\lambda}_{\sigma(x),\sigma(x'),(2)} e^{-iy_t}\right)\right|}{\sum_{x''=1}^{M} q_{\sigma(x),x'',(2)} \left|\exp\left(\tilde{\lambda}_{\sigma(x),x'',(2)} e^{-iy_t}\right)\right|}.\end{aligned} \tag{3.11}$$

Let $x \in \{1, ..., M\}$ *be fixed. Applying (3.11) a first time with* $x' = x$ *and a second time with any* x'*, we get*

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

and so

$$\forall x', \quad \frac{q_{x,x',(1)}}{q_{x,x,(1)}} = \frac{q_{\sigma(x),\sigma(x'),(2)}}{q_{\sigma(x),\sigma(x),(2)}} \tag{3.12}$$

and

$$\forall x', \quad \tilde{\lambda}_{x,x',(1)} - \tilde{\lambda}_{x,x,(1)} = \tilde{\lambda}_{\sigma(x),\sigma(x'),(2)} - \tilde{\lambda}_{\sigma(x),\sigma(x),(2)}. \tag{3.13}$$

Now, since $\sum_{x'} q_{x,x',(1)} = 1 = \sum_{x'} q_{\sigma(x),\sigma(x'),(2)}$*, due to (3.12), 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 (3.13), we get* $\tilde{\lambda}_{x,x,(1)} = \tilde{\lambda}_{\sigma(x),\sigma(x),(2)}$ *and, applying again (3.13), we conclude that*

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

3.4 Numerical results and model comparison

3.4.1 Data

In this work, 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 wind data for the point with geographical coordinates (47.5^0 N, 5^0 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 work 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 (Ailliot and Monbet, 2012) for the wind speed at the same location since the diurnal components can be neglected during the winter season. Following (Ailliot and Monbet, 2012), 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.

3.4.2 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 3.1 and 3.2). 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**$_{(U,\Phi)}$ 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

	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 3.1: 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 3.2: BIC values for the various bivariate models. The last column gives the number of parameters. The bold values correspond to the selected models.

both parameterizations (polar or Cartesian) since the differences in BIC values are relatively small.

3.4.3 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 3.3 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

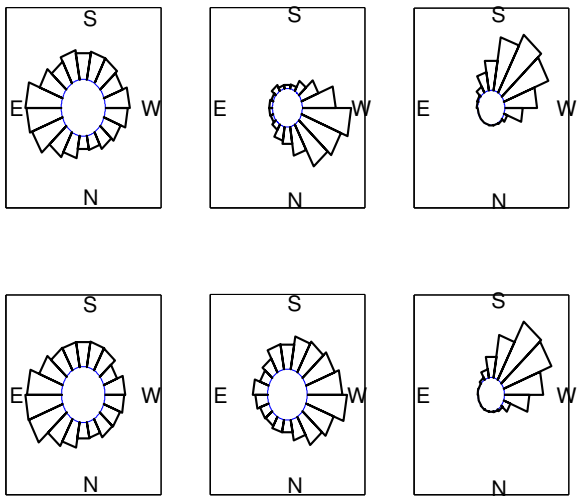

Figure 3.3: Rose plot of the wind direction in the three regimes identified on wind speed by **NHMS-AR**$_{(U,\Phi)}$ model (top) and by **NHMS-AR**$_{(u,v)}$ model (bottom).

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 3.3).

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 Cattiaux et al. (2013), 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 3.3. It shows a clear link between the regimes identified by the **NHMS-AR**$_{(u,v)}$ model and the large-scale regimes. For example the first regime of the **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 **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

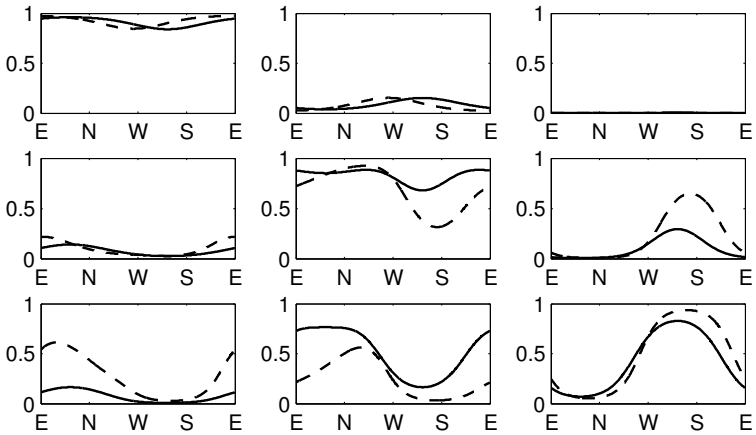

Figure 3.4: Matrix of non-homogeneous transitions of **NHMS-AR**$_{(u,v)}$ model (plain line) and **NHMS-AR**$_{(U,\Phi)}$ model (dashed line)

Regime	BL	AR	NA0-	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 3.3: Joint probability of occurrence of the three regimes identified by the the **NHMS-AR**$_{(u,v)}$ model (lines) and the large-scale regimes provided by J. Cattiaux (see Cattiaux et al. (2013)) in columns.

with the climatology of the area.

Both **NHMS-AR**$_{(u,v)}$ and **NHMS-AR**$_{(U,\Phi)}$ models have similar transition probabilities (see Figure 3.4) with a more pronounced dependence on the wind direction for the **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.

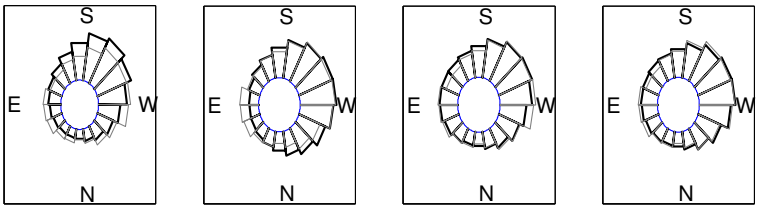

Figure 3.5: Rose plot of the marginal distribution of wind direction for the various models with $M = 4$ (resp. $M = 3$) regimes for Φ_t (resp.$\{u_t, v_t\}$) and order $s = 2$. From left to right: **HMS-AR**$_{(u,v)}$, **HMS-AR**$_{(U,\Phi)}$, **NHMS-AR**$_{(u,v)}$, **NHMS-AR**$_{(U,\Phi)}$

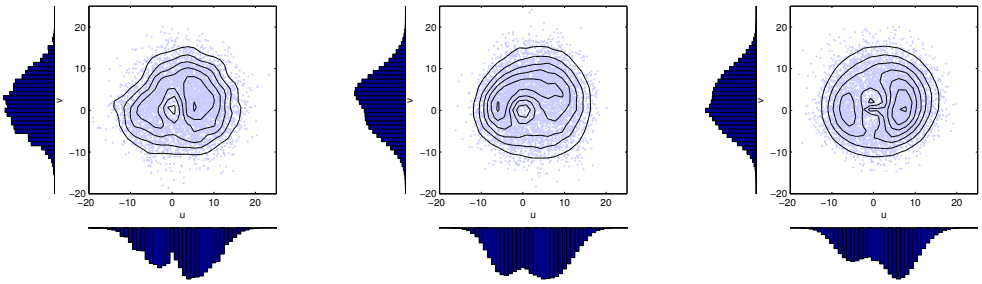

Figure 3.6: Joint distributions of $\{u_t, v_t\}$ for observed wind (left) and wind simulated with **NHMS-AR**$_{(u,v)}$ model (middle) and **NHMS-AR**$_{(U,\Phi)}$ model (right)

3.4.4 Marginal distributions

According to Figure 3.5, 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 **NHMS-AR**$_{(U,\Phi)}$ model seems to perform slightly better than the **NHMS-AR**$_{(u,v)}$ model.

The joint distribution of $\{u_t, v_t\}$ is globally well reproduced by both non-homogeneous models (see Figure 3.6). Simulated data exhibit two modes like in the original data. The modes seem to be slightly better located with the **NHMS-AR**$_{(U,\Phi)}$ model. It may be due to the small differences in the non-homogeneous transition probabilities, see Figure 3.4. The **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 miss-specified. This discrepancy may be reduced by developing alternative estimation methods which would give more importance to the stationary distribution of the process. This will be the topic of future research.

3.4.5 Dependence structure

All the models reproduce approximatively the first lags of autocorrelation function of $\{U_t\}$ (see Figure 3.7) and the circular autocorrelation of Φ (not shown) defined as (see (Fisher and Lee, 1983))

$$\rho(h) = \frac{E[\cos(\Phi_0)\cos(\Phi_h)]+E[\sin(\Phi_0)\sin(\Phi_h)]-E[\sin(\Phi_0)\cos(\Phi_h)]-E[\cos(\Phi_0)\sin(\Phi_h)]}{E[\cos(\Phi_0)^2]E[\sin(\Phi_0)^2]-E[\sin(\Phi_0)\cos(\Phi_0)]^2} \tag{3.14}$$

for any positive integer h. To further validate the models, we have also plotted the various terms which appear in (3.14).

According to Figure 3.8, 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 3.7 shows the cross-correlation function between the time series $\{\cos(\Phi_t)\}$ and $\{\sin(\Phi_t)\}$. The sample cross-correlation function is at its maximum value for a lag between 18 hours and 24 hours, with the time series $\{\sin(\Phi_t)\}$ being in advance of the time series $\{\cos(\Phi_t)\}$. 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 3.9). Note that the complex parametrization of the von Mises autoregressive models (see Section 3.2.4) permits to model rotation in a prevailing direction and significantly improves the boxplot shown on Figure 3.9 compared to models with real parametrization (not shown). One can also remark that the first order autoregressive matrices of the **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 3.9 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 $\{\cos(\Phi_t)\}$ and $\{\sin(\Phi_t)\}$. 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

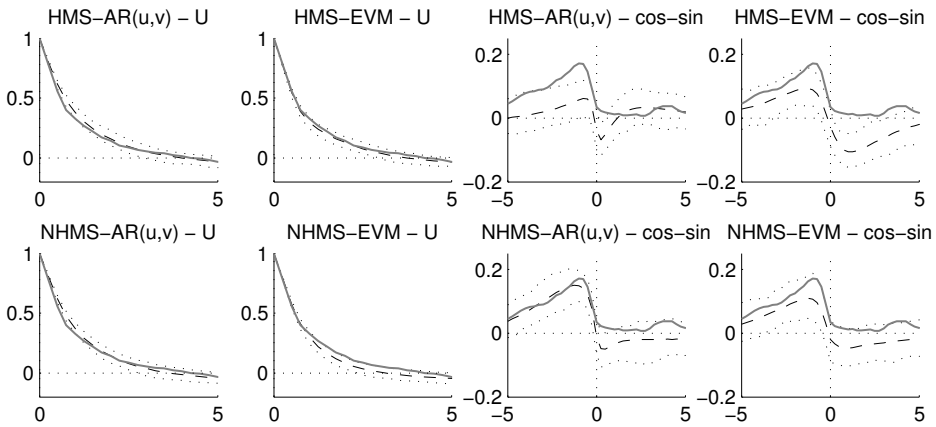

Figure 3.7: Correlation function of $\{U_t\}$ and cross-correlation functions between the time series $\{\cos(\Phi_t)\}$ and $\{\sin(\Phi_t)\}$ for the various models. The full grey line corresponds to the sample functions and the dashed line to the fitted model with a 95% prediction intervals (dotted line). The distributions for the fitted model was obtained by simulation. Time on the x-axis is expressed in days. (a): **HMS-AR**$_{(u,v)}$, (b): **HMS-AR**$_{(U,\Phi)}$, (c): **NHMS-AR**$_{(u,v)}$, (d): **NHMS-AR**$_{(U,\Phi)}$

marginal distribution and thus models of order $s = 2$ seem to provide a good compromise.

3.5 Conclusion

In this work we propose to model bivariate wind time series considering Cartesian In this work 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

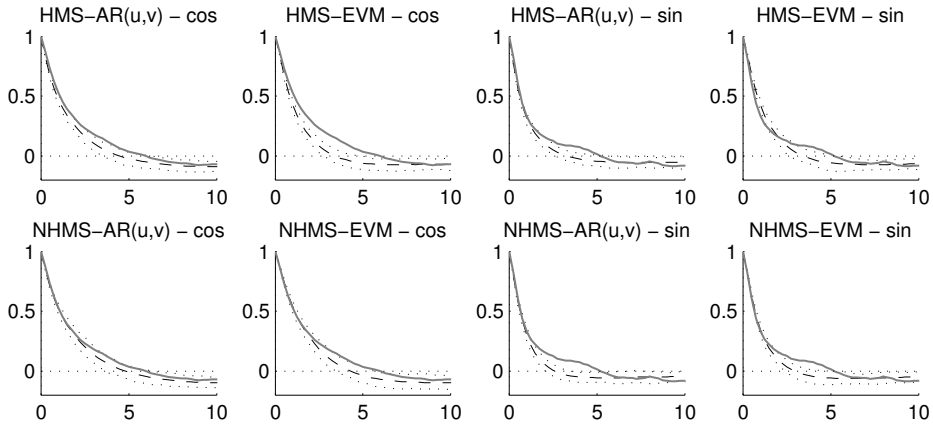

Figure 3.8: Autocorrelation functions of the time series $\{\cos(\Phi_t)\}$ and $\{\sin(\Phi_t)\}$ for the various models. The full grey line corresponds to the sample functions and the dashed line to the fitted model with a 95% prediction intervals (dotted line). The distributions for the fitted model was obtained by simulation. Time on the x-axis is expressed in days. (a): **HMS-AR**$_{(u,v)}$, (b): **HMS-AR**$_{(U,\Phi)}$, (c): **NHMS-AR**$_{(u,v)}$, (d): **NHMS-AR**$_{(U,\Phi)}$

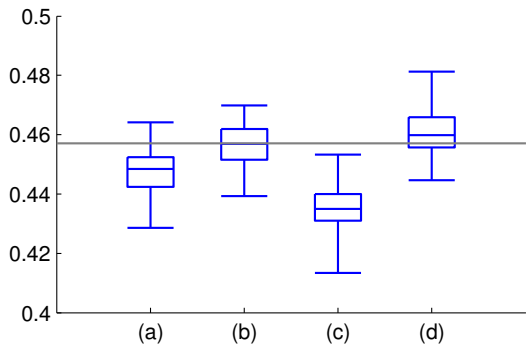

Figure 3.9: Frequency of anticlockwise rotations between successive observations for the various models with $M = 4$ regimes and autoregressive model of order $s = 2$. The grey line corresponds to the value obtained on the data (45.4 % of anticlockwise rotations against 54.6% of clockwise rotations). 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**$_{(U,\Phi)}$, (c): **NHMS-AR**$_{(u,v)}$, (d): **NHMS-AR**$_{(U,\Phi)}$

directions (northeast for anti-cyclonic conditions and southwest 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.

Chapter 4

Markov-Switching AutoRegressive models for Cartesian components of wind fields in the North-East Atlantic

This chapter is the object of an on-going work with P. Ailliot, J. Cattiaux and V. Monbet.

Several multi-site generators of $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$ wind conditions are proposed in this work. A regime-switching framework is introduced to account for the alternation of intensity and variability that is observed on wind conditions due to the weather state. This modeling consists in blocking 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 observed or latent. A hidden Markov-Switching Vector AutoRegressive model is introduced and compared to an unconditional and several conditional Vector AutoRegressive models with observed regime-switching. Various questions are explored such as the modeling of the regime in a multi-site context, the extraction of relevant clusterings from extra-variables or from wind data and the link between regimes extracted from the fitting of the hidden MS-AR model and large-scale weather types derived from a descriptor of atmospheric circulation. The proposed models reproduce the average space-time motions of wind conditions and we show the advantage of regime-switching models in reproducing the alternation of variability of wind conditions.

4.1 Introduction and general context

4.1.1 Introduction

Stochastic weather generators have been used to generate artificial sequences of small-scale meteorological data with statistical properties similar to the one of the dataset that is used to calibrate the model. Various wind conditions generators at a single site have been proposed in the literature. However very few models were introduced in the multi-site context (Haslett and Raftery, 1989). Artificial sequences of wind conditions provided by stochastic weather generators enable to assess related risks in impacts studies, see for instance (Hofmann and Sperstad, 2013). We propose in this work a multi-site generator for Cartesian components of surface wind. As far as we know, only a few models have been proposed to model time series of Cartesian coordinates of wind $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$ (Hering and Genton, 2010; Ailliot et al., 2006b; Wikle et al., 2001; Modlin et al., 2012). These models are mostly purposed for short-term wind prediction and not for the generation of artificial conditions of $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$. They are not focused on reproducing the same statistics we are interested in, that is to say the marginal distribution and the spatio-temporal dynamic.

In the North-East Atlantic, the spatio-temporal dynamic of wind is complex because a regime-switching is observed due to large-scale weather regimes. It induces an alternation between periods with high temporal variability with more stable periods. Introducing regime-switching in the modeling, as it is proposed in this work, permits to reproduce the various temporal dynamics and scales present in the wind data. In this work, we propose various Vector AutoRegressive (VAR) Models with regime-switching. One of the challenges is to propose a regime-switching that is physically consistent and that enables to describe appropriately the local observation by a VAR model.

In practice, blocking a time series into regimes consists in partitioning it into periods of time in which the series is homogeneous and can be described by a single model. Depending on the availability of good descriptors of the current weather state, regime-switching can be achieved through models with an observed or a latent regime-switching. The regime is said to be observed when regimes are identified *a priori* for instance before modeling the local dynamic. In this case, clustering methods are run on adequate variables to obtain relevant regimes. In practice, regimes can be extracted from extra-variables, such as descriptors of atmospheric circulation (see for instance (Bardossy and Plate, 1992; Wilson et al., 1992)), or from local variables. For instance, separation of dry-wet states has been widely used to derive observed regimes when various meteorological variables are considered, see (Richardson, 1981; Flecher et al., 2010). When considering wind models, wind direction can be accounted for since it is a good descriptor of synoptic conditions. In (Gneiting et al., 2006), wind direction is used both to extract regimes and in the parametriza-

tion of the predictive distribution. *A priori* clusterings based on global or local variables are discussed in Section 4.4.

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 Model have been widely used for meteorological data (Zucchini and Guttorp, 1991; Hughes et al., 1999; Thompson et al., 2007). Markov-Switching AutoRegressive (MS-AR) models appear as a generalization of Hidden Markov Models (HMM) in allowing temporal dynamics within the regimes (Hamilton, 1989). Models with regime-switching improve the modeling of wind intensity time series compared with classical AutoRegressive-Moving Average (ARMA) models, see (Ailliot and Monbet, 2012) where wind speed is modeled at one site. MS-AR models are introduced in section 4.2 and their inference is described.

In the multi-site context, the regime can be regional for all sites and remains scalar (Ailliot et al., 2009) or it can be introduced as a site-specific regime (Wilks, 1998; Kleiber et al., 2012; Khalili et al., 2007; Thompson et al., 2007), which enables to account for a wide range of space-time dependence. However a site-specific regime appears to be computationally challenging (Wilks, 1998). A comparison of MS-AR models with site-specific regimes against regional regime is detailed Section 4.3.

4.1.2 Wind data

The data under study are west-east and north-south surface wind components $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$ at 10 meters above sea level extracted from the ERA Interim Full dataset produced by the European Center of Medium-range Weather Forecast (ECMWF). It can be freely downloaded and used for scientific purposes at the URL `http://data.ecmwf.int/data/`.

We focus on gridded locations between latitudes 46.5°N and 48°N and longitudes 6.75°W and 10.5°W (see Figure 4.1). The dataset we have extracted consists of 32 blocks of months December and January of wind data from December 1979 to January 2011. 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. In order to study the spatial coherence of a common regime to all the locations, an homogeneous area is sought. A spatial hierarchical clustering has been realized to choose an homogeneous area (see Figure 4.1). The clustering is run on the process of moving variance of wind speed, which is described more precisely in Section 4.5. This process is a good descriptor of the alternate of temporal variability of time series (see Figure 4.2) and it is computed as the variance of wind speed over nine consecutive time. The dendogram suggests the use of four clusters

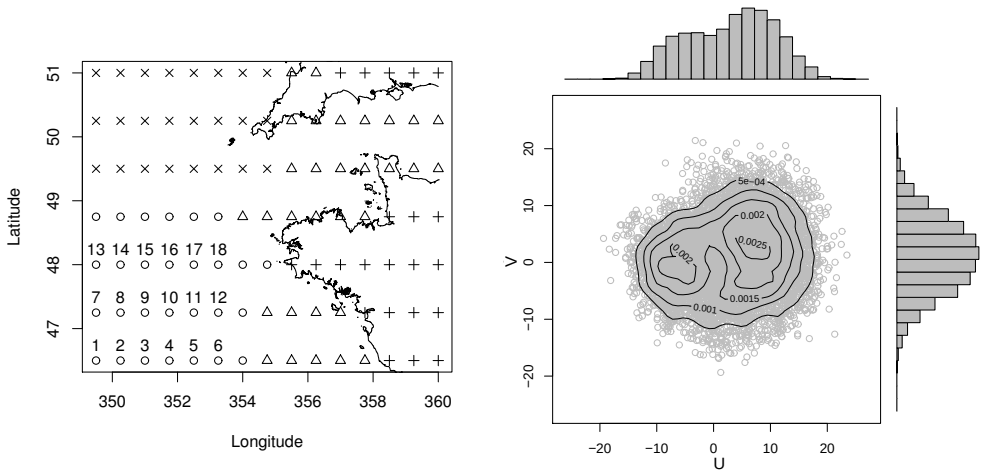

Figure 4.1: Left: Spatial hierarchical clustering of the moving variance associated to the process of wind speed with four clusters. Right: Joint and marginal distribution of $\{u_t, v_t\}$ at the central location 10, contour lines of the estimated joint density.

that are depicted on Figure 4.1. These four clusters are likely to be divided in an inland cluster, an intermediate cluster between ocean and land, a cluster corresponding to storms that propagate into the Bay of Biscay and another for storms that propagate toward northern Europe.

Components $\{u_t\}$ and $\{v_t\}$ admit a complex relationship, the joint distribution and the cross-correlation of $\{u_t, v_t\}$ reflect a part of this complexity (Figure 4.1). The margin of $\{u_t\}$ reveals two separate modes whereas the one of $\{v_t\}$ does not exhibit a clear bi-modality. The very few points around the point (0,0) indicate 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 transformation is used on both components $\{u_t\}$ and $\{v_t\}$. This transformation with $\alpha > 1$ aims at filling the hole around $(0, 0)$ in order to facilitate the modeling of the bi-modality

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

where $\{U_t\}$ and $\{\Phi_t\}$ respectively denote wind speed and wind direction. In practice, α is chosen empirically equal to 1.5.

4.2 Markov-Switching AutoRegressive models

4.2.1 The models

In the present work, 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 models described here are Vector AutoRegressive models. Given the current value of S_t, the observation $\boldsymbol{Y}_t$ is written as:

$$\boldsymbol{Y}_t = \boldsymbol{A}_0^{(S_t)} + \boldsymbol{A}_1^{(S_t)}\boldsymbol{Y}_{t-1} + \boldsymbol{A}_2^{(S_t)}\boldsymbol{Y}_{t-2} + ... + \boldsymbol{A}_p^{(S_t)}\boldsymbol{Y}_{t-p} + (\boldsymbol{\Sigma}^{(S_t)})^{1/2}\boldsymbol{\epsilon}_t, \quad (4.1)$$

$\boldsymbol{Y}$ represents the observed power-transformed K-dimensional process. For $i \in \{1, ..., M\}$, $\boldsymbol{A}_0^{(i)}$ is a K-dimensional vector, $\boldsymbol{A}_1^{(i)}, ..., \boldsymbol{A}_p^{(i)}, \boldsymbol{\Sigma}^{(i)}$ are $K \times K$-matrices and $\boldsymbol{\epsilon}$ is a Gaussian white noise of dimension K. In the sequel, K equals 36 when the dataset of 18 locations is considered. Conditional independences 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):

$$\begin{tikzcd}
\cdots \arrow[r] \arrow[d] & S_{t-1} \arrow[r] \arrow[d] & S_t \arrow[r] \arrow[d] & S_{t+1} \arrow[r] \arrow[d] & \cdots \arrow[d] \\
\cdots \arrow[r] & \boldsymbol{Y}_{t-1} \arrow[r] & \boldsymbol{Y}_t \arrow[r] & \boldsymbol{Y}_{t+1} \arrow[r] & \cdots
\end{tikzcd}$$

In the above model, the regime S can be latent or observed, both cases are discussed respectively in Sections 4.3 and 4.4. In both cases, the parameters estimation of the model can be performed by maximum likelihood but in a different way in each case.

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 et al., 2014) and in (Vrac et al., 2007) when regimes are defined *a priori*). In the context of multi-site models, the choice of the covariate of non-homogeneous transitions is delicate, we do not discuss this topic here and only consider homogeneous models.

To avoid over-parameterization of the conditional models, we work in a first step with a reduced dataset. In the following all the proposed models will be fitted on the set of the sites (1,6,10,13,18).

4.2.2 Estimation by maximum likelihood

Firstly, let 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 4.4. Assume that $\boldsymbol{y}_{-1}$ and $\boldsymbol{y}_0$ are

observed, the complete log-likelihood, associated to an autoregressive order $p = 2$, is written as:

$$\begin{aligned}\log(\mathcal{L}(\theta; \boldsymbol{y}_1, ...\boldsymbol{y}_T, s_1, ...s_T|\boldsymbol{y}_{-1}, \boldsymbol{y}_0)) &= \log(\mathcal{L}(\theta^{(\boldsymbol{Y})}; \boldsymbol{y}_1^T|\boldsymbol{y}_{-1}, \boldsymbol{y}_0, s_1^T)) \\ &\quad + \log(\mathcal{L}(\theta^{(S)}; s_1^T|\boldsymbol{y}_{-1}, \boldsymbol{y}_0)), \quad (4.2)\end{aligned}$$

with $\theta = (\theta^{(S)}, \theta^{(\boldsymbol{Y})})$, where $\theta^{(\boldsymbol{Y})}$ 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, which is a stochastic matrix, and $\boldsymbol{y}_1^T = (\boldsymbol{y}_1, ..., \boldsymbol{y}_T)$. Let denote $n_{i,j}$ the number of occurrences of the event $\{(S_t, S_{t+1}) = (i, j)\}$ for all $t \in \{1, ..., T-1\}$, $n_{i,.} = \sum_{j=1}^M n_{i,j}$ and $n_i = n_{i,.} + \delta_{\{s_T=i\}}$, where δ is the Kronecker symbol, the total number of occurrence of the regime i.

$$\begin{aligned}&\log(\mathcal{L}(\theta^{(\boldsymbol{Y})}; \boldsymbol{y}_1, ..., \boldsymbol{y}_T|\boldsymbol{y}_{-1}, \boldsymbol{y}_0, s_1^T)) \\ =& \sum_{t=1}^T \log(p(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}, \boldsymbol{y}_{t-2}, s_t)) \text{ by Markovian properties} \\ =& \sum_{i=1}^M \sum_{t\in\{t|s_t=i\}} \log(p(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}, \boldsymbol{y}_{t-2}, s_t)) \\ =& \sum_{i=1}^M n_i(-\tfrac{d}{2}\log(2\pi) - \tfrac{1}{2}\log(\det(\boldsymbol{\Sigma}^{(i)})) - \sum_{t\in\{t|s_t=i\}} \tfrac{1}{2}\boldsymbol{e}_t'(\boldsymbol{\Sigma}^{(i)})^{-1}\boldsymbol{e}_t,\end{aligned}$$

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

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

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 computation of the optimal estimates of $\boldsymbol{A}_1^{(i)}$ and $\boldsymbol{A}_2^{(i)}$ is performed via the writing of 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 write $\boldsymbol{A}^{(i)} = \begin{pmatrix} \boldsymbol{A}_1^{(i)} & \boldsymbol{A}_2^{(i)} \\ \mathbf{Id}_{\boldsymbol{K}} & \mathbf{0} \end{pmatrix}$ 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

$$\hat{\boldsymbol{A}^{(i)}} = \Big(\sum_{t\in\{t|s_t=i\}} \boldsymbol{Z}_t\boldsymbol{Z}_{t-1}'\Big)\Big(\sum_{t\in\{t|s_t=i\}} \boldsymbol{Z}_{t-1}\boldsymbol{Z}_{t-1}'\Big)^{-1}. \quad (4.3)$$

The other optimal estimates are:

$$\hat{\boldsymbol{A}}_0^{(i)} = (\mathbf{Id}_{\boldsymbol{K}} - \hat{\boldsymbol{A}}_1^{(i)} - \hat{\boldsymbol{A}}_2^{(i)})\hat{\boldsymbol{\mu}}^{(i)} \tag{4.4}$$

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

$$\hat{\boldsymbol{\Sigma}}^{(i)} = \frac{1}{n_i} \sum_{t \in \{t|s_t=i\}} \hat{\boldsymbol{e}}_t \hat{\boldsymbol{e}}_t^{'}, \tag{4.5}$$

$\hat{\boldsymbol{\Sigma}}^{(i)}$ 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:

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

the associated maximum likelihood estimator is

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

notice that $\hat{\theta}^{(S)}$ satisfies naturally the constraint of a stochastic matrix.

When only observations of the process $\boldsymbol{Y}$ are available and the realizations of S are not given *a priori*, like in Section 4.3, 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 complete log-likelihood function based on the observations $\boldsymbol{Y}$:

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

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 proceeds into cycling through two steps: the Expectation-step and the Maximization-step (Wu, 1983; Dempster et al., 1977). The E-step is performed through Forward-Backward recursions (see (Hamilton, 1990) for hidden MS-AR models) that enable to compute the smoothing probabilities $P(S_t|\boldsymbol{Y}_1^T = \boldsymbol{y}_1^T)$. At M-step, optimal expressions of parameters of $\theta^{(\boldsymbol{Y})}$, given in (4.3), (4.4) and (4.5), are used. However in each regime i, each observation $\boldsymbol{y}_t$ is weighted by the probability $P(S_t = i|\boldsymbol{Y}_1^T = \boldsymbol{y}_1^T)$, for instance

$$\hat{\boldsymbol{\mu}}^{(i)} = \frac{1}{\sum_{t=1}^T P(S_t = i|\boldsymbol{Y}_1^T = \boldsymbol{y}_1^T)} \sum_{t=1}^T P(S_t = i|\boldsymbol{Y}_1^T = \boldsymbol{y}_1^T)\boldsymbol{y}_t.$$

The estimate of the transition matrix are obtained from quantities $P(\{S_t = i, S_{t+1} = j\}|\boldsymbol{Y}_1^T = \boldsymbol{y}_1^T)$ that are derived at the E-step.

The following notations are used in the sequel: AP-MS-VAR$_{\mathrm{C}}$ to denote the *a priori* regime-switching model associated to the clustering C and H-MS-VAR to denote the hidden regime-switching model.

4.3 From a single-site to a multi-site hidden MS-AR model

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 of the literature, see (Zucchini and MacDonald, 2009) for a wide range of applications of Hidden Markov Models. In (Ailliot et al., 2014), a single-site model is proposed to model $\{u_t, v_t\}$, the proposed hidden Markov-Switching AutoRegressive model reveals good qualities to describe marginal and joint distribution of $\{u_t, v_t\}$ as well as the temporal dynamics of the wind at one location.

In this section, the assumption of a common regional regime is investigated, we show that this assumption is acceptable. The homogeneous MS-AR model introduced in (Ailliot et al., 2014) for $\{u_t, v_t\}$ with $M = 3$ regimes and an autoregressive order $p = 2$ has been fitted at each site. Most likely regimes associated to the data are extracted from the estimation procedure of hidden MS-AR models, at each time the regime is $\underset{S_t \in \{1,\cdots,M\}}{\arg\max}\ P(S_t|\boldsymbol{Y}_1^T = \boldsymbol{y}_1^T)$. The spatio-temporal coherence of the regimes of each of the 18 sites is checked and reveals a strong homogeneity that urges to use a regional regime in this area.

In order to compare properly the regimes, they are ordered according to the increasing value of the determinant of $\boldsymbol{\Sigma}^{(i)}$. 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). The two other regimes correspond to cyclonic conditions with westerly wind and higher temporal variability in the intensity (see Figure 4.2). 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 In Figures 4.2, we can notice that stable wind conditions observed in the first regime are associated to weak values of the moving mean and variance process whereas cyclonic conditions and volatile periods observed in the second and third regimes are characterized by higher values of moving mean and variance.

Coefficients of the AutoRegressive process $\boldsymbol{Y}$ in each regime and the transi-

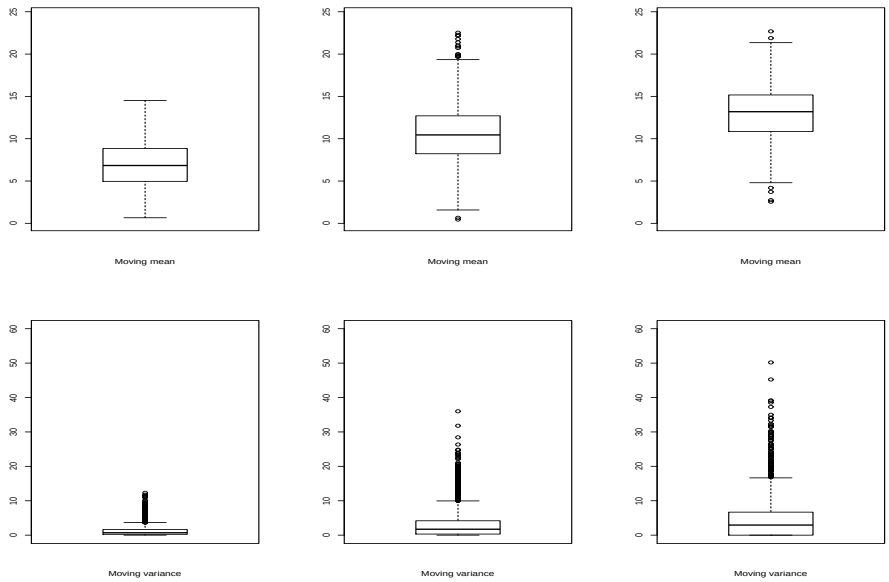

Figure 4.2: Top line: moving means computed on two days intervals in each regime of the H-MS-VAR model fitted at site 10. Bottom line: moving mean square error of wind speed around its moving mean, in each regime. Moving means and errors are computed over an interval of two days .

	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

Table 4.1: Diagonal of the transition matrix $\mathbf{\Pi}$ at each site, coefficients of the AutoRegressive model in each regime at each site and logarithm of the determinant of $\mathbf{\Sigma}^{(i)}$.

tion matrix at each site are very comparable and spatially coherent (see Table 4.1). 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.

Moreover, the sequences of regimes can be compared in Figure 4.3. Time series of *a posteriori* regimes and wind speed are depicted. Homogeneity is strong at the three locations which suggests the use of a regional regime. The two last regimes are the less coherent from one site to another, this is partly explained by the fact that these regimes are the less stable especially the third one (see Table 4.1). Moreover, we can notice propagations of eastward wind events, indeed the darkest regimes are often first observed at western stations (station 1) and then observed at more eastern sites (10 and 18). The bottom panel of the Figure 4.3, which depicts the sequences of regimes associated with the model fitted on the all set of locations with a common regime to all locations, reveals that this regional regime is coherent with the local ones although it is less persistent.

In Figure 4.4, probability of occurrence of a regime at a given location conditionally on the same regime that occurs at the same time at site 10, are depicted. On each picture, conditional probabilities should be compared to the reference value given at location 10, which is 1. The first regime has the best spatial coherence and the third regime, which is the less persistent regime, is the less coherent spatially. The assumption of a regional regime seems appropriate and is then kept for the modeling of the multi-site wind in the following.

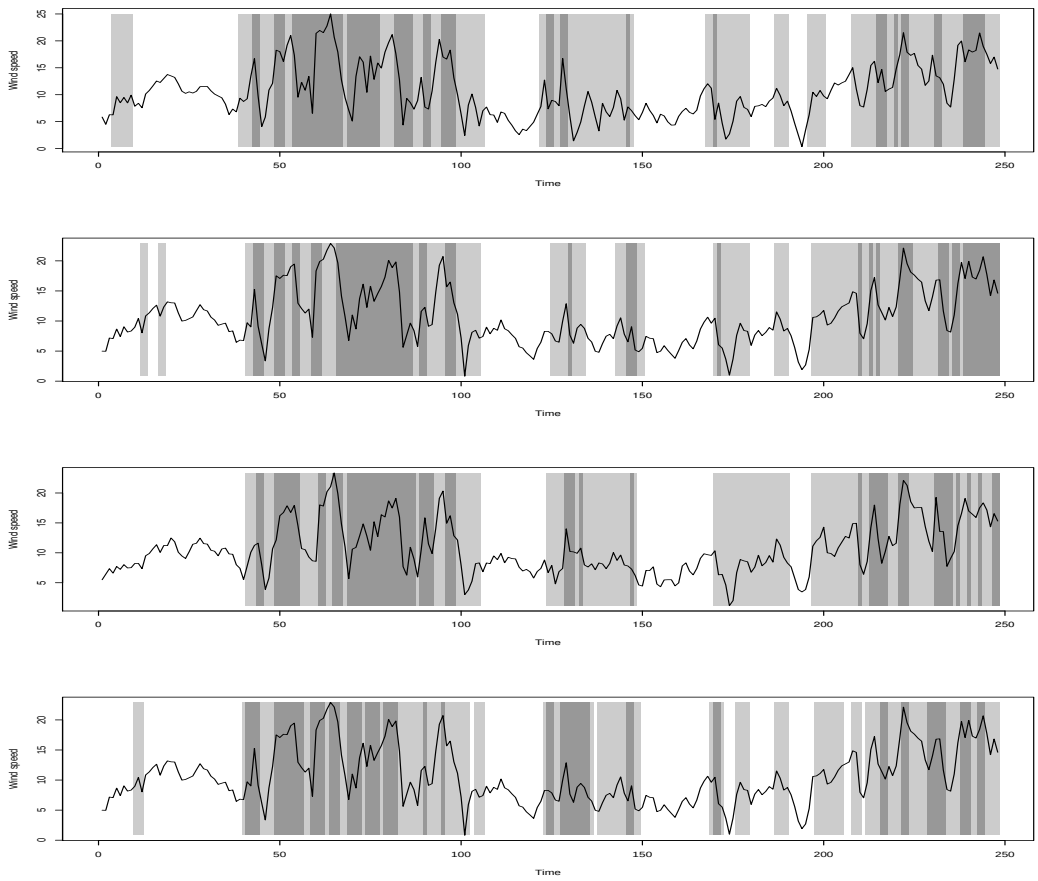

Figure 4.3: Time series of wind speed and *a posteriori* regimes from the fitting of a MS-AR. The lighter is the grey, the smaller is the determinant of $\mathbf{\Sigma}^{(i)}$. From top to bottom: sites 1, 10 and 18 when the model is fitted at a single location, bottom line: extracted regimes when the model is fitted at the 5 locations (1,6,10,13,18).

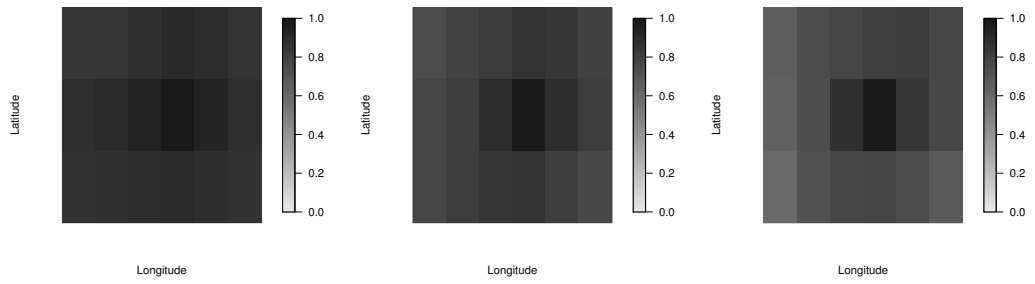

Figure 4.4: Conditional probabilities of occurrence of regime $i = 1, \cdots, 3$ at one site conditionally on the occurrence of the same regime at site 10.

4.4 Observed regime-switching AutoRegressive models

Conversely to the previous section, one may derive the regime separately from the fitting of the conditional model. When considering *a priori* regime-switching, the derivation of regimes can be done with appropriate clustering methods. We seek for weather states that are distinct one from the other and in which the data are homogeneous. We expect here to extract, among others, a regime with westerly conditions and another with easterly conditions. Clustering can be run on the local studied variables or on extra variables. When it is processed on the local data, the weather states might be more appropriate to the data but may contain fewer information about the synoptic scale. In the second case, more information about the synoptic scale is available which may improve the meteorological consistency of the regimes. In this subsection, we propose three clusterings, which differ by the choice of the clustering method and by the choice of the descriptors of the variables to derive the *a priori* regimes.

4.4.1 Derivation of observed regimes from extra-variables

The large-scale weather regimes commonly used in climate studies is considered as a first clustering. The four wintertime weather regimes of (Cattiaux et al., 2013) are obtained over the North-Atlantic / European sector (90W-30E / 20-80N) by a kmeans-clustering of the time series associated to the first Empirical Orthogonal Functions (EOF) of the 3607 daily maps of 500 mb geopotential anomalies (mean-corrected fields, days of December, January and February 1981–2010). In winter, four regimes are identified and described in various references (Michelangeli et al., 1995; Cassou, 2008; Najac, 2008). They correspond to the two phases of North-Atlantic Oscillation (NAO+ and

NAO-), the blocking (BL) and the Atlantic Ridge (AR). In France in wintertime, these four regimes respectively correspond to privileged flows that are respectively: south-western flows (NAO+), western flows (NAO-), southern or eastern stable flows (BL) and northern flows (AR). Let denote C_{Z500} this clustering.

4.4.2 Derivation of observed regimes from the local variables

C_{Z500} provides persistent regimes in which the estimation and simulation of the conditional model lead to a satisfying description of $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$. However when using a k-means clustering on descriptors of the local wind, regimes are not persistent enough to estimate reliably the conditional VAR model. Consequently, in this subsection, we perform the clustering via a Hidden Markov Model with Gaussian probability of emission. The hidden structure of Markov chain provides more stable regimes than with a k-means clustering. The EM-algorithm is used to process the clustering and the number of regimes is chosen at three. This number of clusters provides the most physically relevant regimes. Two kinds of descriptors of the data are proposed.

The first partition of the data is obtained by clustering the time series associated to the first EOF of the anomalies of $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$. These time series correspond to the projections of the anomalies on the EOFs, which are eigenvectors of the spatial covariance matrix of the anomalies time series. This decomposition enables to extract the main modes of variability of the spatio-temporal process. In practice, the two first EOF, which explain 94% of the total variance, are kept. Let denote respectively $\mathrm{C}_{EOF-(u,v)}$ this clustering.

Furthermore, in order to find a clustering that may be better adapted to the description of the conditional distribution by an autoregressive model, we consider a method that involves descriptors of the conditional distribution of $p(\boldsymbol{Y}_t|\boldsymbol{Y}_{t-1})$. A simplified way to account for such descriptors 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 to construct regimes that discriminate well the variances of the process $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$. Let denote $\mathrm{C}_{Diff(u,v)}$ the associated clustering.

4.4.3 Comparison and selection

The proposed clusterings are compared through various quantitative and qualitative criteria in order to select the clustering that is the most physically meaningful and appropriate in terms of conditional autoregressive models. For a proper comparison, except for C_{Z500}, in each clustering the regimes are ordered according to the determinant of the matrices $\boldsymbol{\Sigma}^{(i)}$. For C_{Z500}, the ranking provided by the determinant and the one provided by the trace of $\boldsymbol{\Sigma}^{(i)}$ differ, this latter ranking is the most physically consistent since NAO+ corresponds

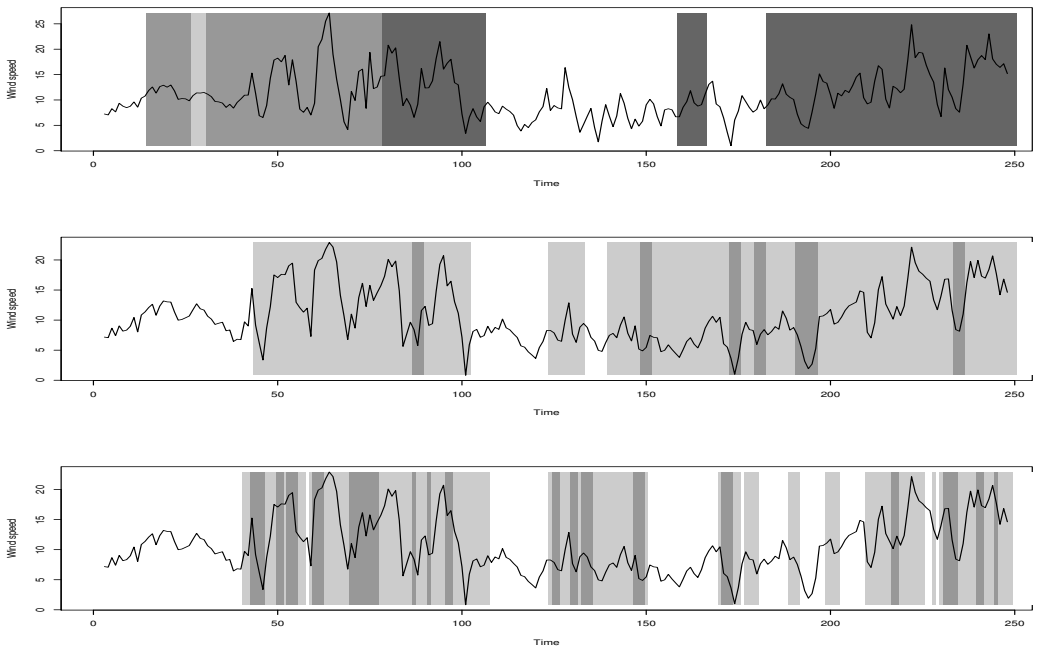

Figure 4.5: Time series of wind speed and *a priori* regimes extracted from the proposed methods above. The darker is the grey, the smaller is the determinant of $\mathbf{\Sigma}^{(i)}$. From top to bottom: C_{Z500}, $\mathrm{C}_{EOF-(u,v)}$ and $\mathrm{C}_{Diff(u,v)}$.

to the most unstable conditions. In the following, the regimes of C_{Z500} are ordered in the following way: BL, AR, NAO-, NAO+.

Sequences of regimes from these three clusterings are shown in Figure 4.5. C_{Z500} has very persistent regimes. In the top panel, one can see that the most stable wind conditions are associated to the BL and AR phases, whereas the most variable wind conditions occur during the two NAO phases. In the middle panel, the second and third regime are generally associated with moderate and high intensity of wind whereas the first regime seems to be associated with stable conditions. In the $\mathrm{C}_{Diff(u,v)}$ clustering (bottom panel), the regimes are less persistent, due to the choice of descriptors that are less stable. Globally the first regime is associated with the most stable conditions and the third one seems to be associated to the most volatile conditions.

On Figure 4.6, the average fields corresponding to each regime of the three clusterings are plotted. The three clusterings enable to extract easterly and westerly regimes. Two regimes of C_{Z500}, the Blocage and NAO-, are not clearly discriminated by the clustering. This is probably due to the lack of local information in the clustering. This lack of discrimination at small scale by a large-scale clustering was also observed in (Najac, 2008). The AR and NAO+

regimes are consistent with the description given in Subsection 4.4.1. Since different descriptors are used, $\mathrm{C}_{Diff(u,v)}$ and $\mathrm{C}_{EOF-(u,v)}$ lead to very different results. $\mathrm{C}_{EOF-(u,v)}$ leads to the most physically consistent regimes: a north-easterly regime, a northwesterly and a southwesterly one, which are flows corresponding to several of the large-scale weather regimes. The two last regimes are associated with strong intensities. From the derivation of this clustering, it is natural to find regimes that correspond to the main mean patterns of variability of the fields. The regimes of $\mathrm{C}_{Diff(u,v)}$ have less straightforward meteorological interpretation. The first regime corresponds to periods of weak intensities. The two last regimes are southwesterly regimes with different intensity from one to the other. These two regimes are not clearly distinct.

The optimal value of the complete log-likelihood of the model is generally a good measure of the relevance of a model. The complete log-likelihood, given in (4.2), evaluated at the maximum likelihood estimator of $\hat{\theta}$ is written in the case of observed shifts as the sum of the two following terms:

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

and

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

Let notice that the maximal log-likelihood of $\theta^{(\boldsymbol{Y})}$ is characterized by the total time spent in each regime and the associated determinant of covariance matrix of innovation (notice that the one-step ahead error of forecast is linked to this quantity). The longer time is spent in a regime with a weak determinant of covariance of innovation, the greater is the log-likelihood (see Table 4.4.3). 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 and of predictability. As a clustering measure, the weaker is the entropy within a cluster, the better is the homogeneity within the cluster. In prediction, the weaker is the entropy, the stronger is the predictability of S_t given S_{t-1}. More generally one tends to minimize this measure. Due to the range of values of the log-likelihood of $\theta^{(\boldsymbol{Y})}$, the value of the one 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 is then ignored. BIC indexes are also given in Table 4.4.3, 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. BIC index enables to reflect the compromise between

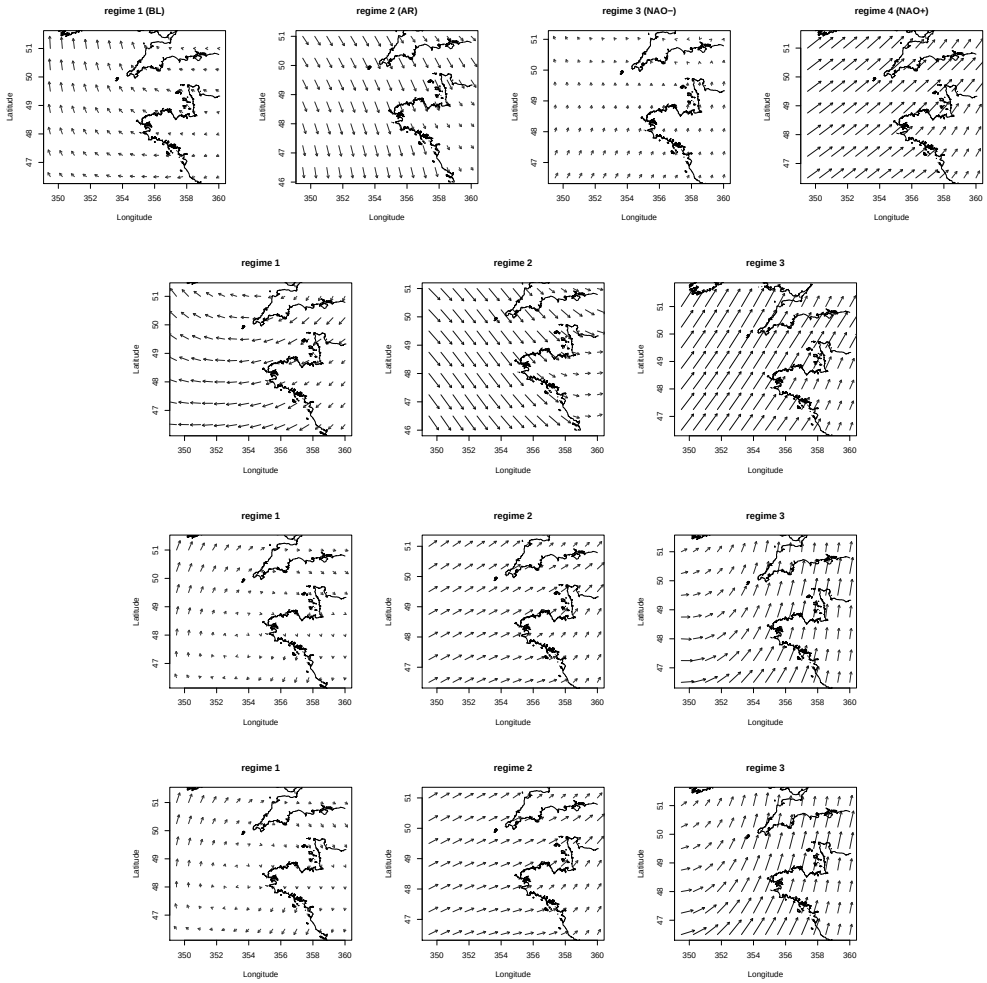

Figure 4.6: Average fields of $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$ for the clusterings, from top to bottom: C_{Z500}, $\mathrm{C}_{EOF-(u,v)}$, $\mathrm{C}_{Diff(u,v)}$ and from the fitting of H-MS-VAR on the set of 5 locations.

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 4.2: 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).

a model with a high likelihood and its parsimony. Notice that one can not compare BIC indexes of *a priori* and latent regime-switching models. However BIC indexes of these two classes of models can be compared to the one 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 since variances of innovation are well separated. The three proposed AP-MS-VAR models lead to a satisfying description of the marginal and joint distribution and space-time covariances (not shown). However 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 4.5. Besides in terms of BIC indexes, the smallest value amongst AP-MS-VAR models is the one of AP-MS-VAR$_{C_{Diff(u,v)}}$ and it is also greater than the one of the VAR model. In the sequel, the VAR model with shifts defined by $C_{Diff(u,v)}$ is kept for further comparisons in simulation.

4.5 Comparison of the multi-site wind models

In this section, we firstly investigate the physical interpretation of the regimes provided by H-MS-VAR. The link between large-scale regimes and the other proposed clusterings is also explored. We finish by a comparison between models VAR(2), AP-MS-VAR$_{C_{Diff(u,v)}}$ and H-MS-VAR in terms of reproducing various scales of the spatio-temporal variability. Especially we focus on the description of the alternate of periods with different temporal variability of wind conditions, we highlight the benefit of using appropriate regime-switching in reproducing this alternate.

A H-MS-VAR model has been fitted to the data with $M = 4$ regimes and parameters of AP-MS-VAR$_{C_{Z500}}$ as initial conditions. However this model do not improve results in simulation and lead to non-interpretable *a posterior* regimes. In the following we consider a H-MS-VAR model with $M = 3$

	$C_{EOF-(u,v)}$					$C_{Diff(u,v)}$					H-MS-VAR				
	BL	AR	NA0 -	NAO+	Total	BL	AR	NA0 -	NAO+	Total	BL	AR	NA0 -	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

Table 4.3: Joint probability of occurrence of the three regimes identified by the proposed models in lines and the large-scale regimes in columns

regimes. Notice that the initialization of the EM-algorithm is robust to initial conditions, initializations with parameters of AP-MS-VAR$_{C_{Diff(u,v)}}$ and AP-MS-VAR$_{C_{EOF-(u,v)}}$ lead to similar results in simulation.

4.5.1 Regimes extracted from hidden MS-VAR model

In Figure 4.6, mean fields of $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$ in each regime extracted from the fitting of the model H-MS-VAR are depicted. These three regimes are associated with three distinct situations. In terms of mean fields, these regimes are similar to the ones of $C_{Diff(u,v)}$. However when comparing Figures 4.3 and 4.5, the sequences of instantaneous behaviors of the regimes differ. Indeed the first regime is associated with the most stable conditions with weak intensity whereas the third one the most variable and stronger winds. The third regime associated with the greatest determinant of $\boldsymbol{\Sigma}^{(i)}$ is the less stable regime and seems to be associated to stormy conditions. The bottom panel of the Figure 4.3 reveals, amongst other things, that the second regime is a precursor to the third one and that this second regime is most of the time associated with raises of wind speed intensity.

A look at the coefficients of the model confirms that the first regime is associated with steady conditions (high AutoRegressive coefficients and weak variance of innovation). This regime is associated with the weakest values of wind speed whereas the two others are related with greater wind intensity especially the third one.

4.5.2 Link between large-scale weather regimes and the other regimes

Meteorology of Europe is mainly driven by the alternate of the characteristic patterns of atmospheric circulation of C_{Z500}. The regimes described by C_{Z500} are related to a large spatio-temporal scale whereas the ones from the hidden MS-VAR are related to a smaller spatio-temporal scale. We explore here the joint occurrences of local regimes provided by $C_{EOF-(u,v)}$, by $C_{Diff(u,v)}$ and by the model H-MS-VAR. For that, the joint probability of occurrence of the regimes identified by the proposed models and the large-scale regimes are com-

puted (Table 4.3). For the three clusterings, the small-scale regimes seem to appear in privileged large-scale weather regimes. The regimes of H-MS-VAR and of $\mathrm{C}_{Diff(u,v)}$ are less persistent than the ones of $\mathrm{C}_{EOF-(u,v)}$ which may explain that the joint occurrences are weaker. Besides the regimes of $\mathrm{C}_{Diff(u,v)}$ and of the H-MS-VAR were already more difficult to interpret than the others. As said previously, the regimes of H-MS-VAR are mainly driven by the conditional autoregressive model in the sense of the likelihood, which may result in a more difficult interpretation. However the regimes of $\mathrm{C}_{EOF-(u,v)}$, which already were easier to interpret, exhibit the strongest link with the large-scale regimes.

4.5.3 Comparison of the MS-VAR models

In this subsection we study the ability of the proposed models to reproduce, via artificial sequences, the statistical properties of the data under study. $N = 100$ sequences of the length of the data are generated with the fitted models and several statistics are computed on these data.

In a first time, marginal statistics at the central site 10 are investigated (see Figure 4.7). Comparing Figures 4.1 and 4.7, one can notice that the distribution of $\{u_t\}$ is well reproduced by the model H-MS-VAR however the one of $\{v_t\}$ is less accurately described. Description of this distribution by AP-MS-VAR$_{\mathrm{C}_{Diff(u,v)}}$ 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, 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 very satisfyingly the general shape of this function and especially the non-separable and anisotropic patterns, see Figure 4.8. The non-separability which is reflected in the asymmetry around the vertical axis at lag 0 is captured by the proposed models.

In order to study further patterns than the average ones, we focus on the ability of the models to reproduce the alternate of temporal variability. Indeed the alternation of different weather states induces an alternation in the intensity and temporal variability of wind. Moving average and moving mean square error around the moving mean have been computed over nine consecutive values of the process $\{u_t, v_t\}$, which corresponds to two days.

In Figure 4.9, moving mean square error of wind speed around its moving

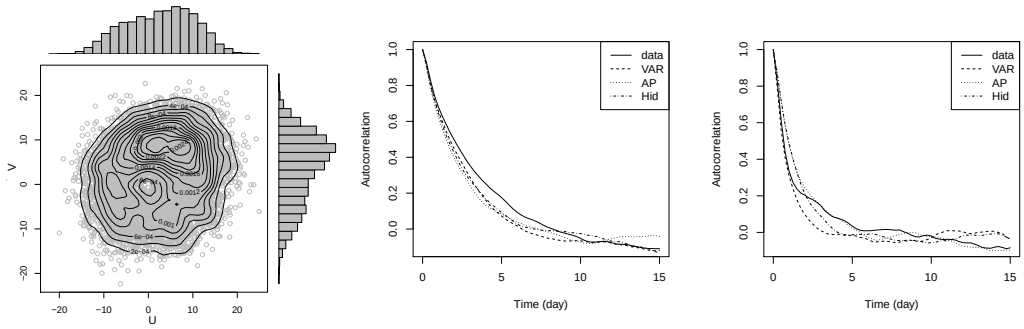

Figure 4.7: Left: joint and marginal distribution of simulated data at site 10 from the model H-MS-VAR. Central and right panels: autocorrelation functions of $\{u_t\}$ and $\{v_t\}$ at site 10 for the reference data, simulated data from the VAR(2), AP-MS-VAR$_{C_{Diff(u,v)}}$ and H-MS-VAR models.

mean at the central site 10 is depicted against its moving mean. Observations reveal a higher variability when the intensity is high but high variability may be associated with weaker values when the moving window overlaps transition time. Models with regime-switching enable to reproduce more temporal variability associated with moderate and high intensity of wind, which is not captured by an unconditional VAR model. We observe that the regime-switching models enable to reproduce high variance intervals associated with high intensity of wind and also the high variability around 5 and 10 $m.s^{-1}$ which corresponds to transitions between weather states. This is ensured by the alternate, driven by a Markov chain, of periods associated with different parameters of the conditional model.

Besides, similar figures than Figure 4.2 indicate that the distributions of the moving variance 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 a characteristic behaviors of these two simulated processes within each regime. Moreover the behavior in each simulated regime is very close to the observed one.

4.6 Discussions and perspectives

In Section 4.3, we study site-specific regimes against regional regime. 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 procedures of likelihood-ratio test, to quantify more precisely to which extent the assumption of a regional regime against a site-

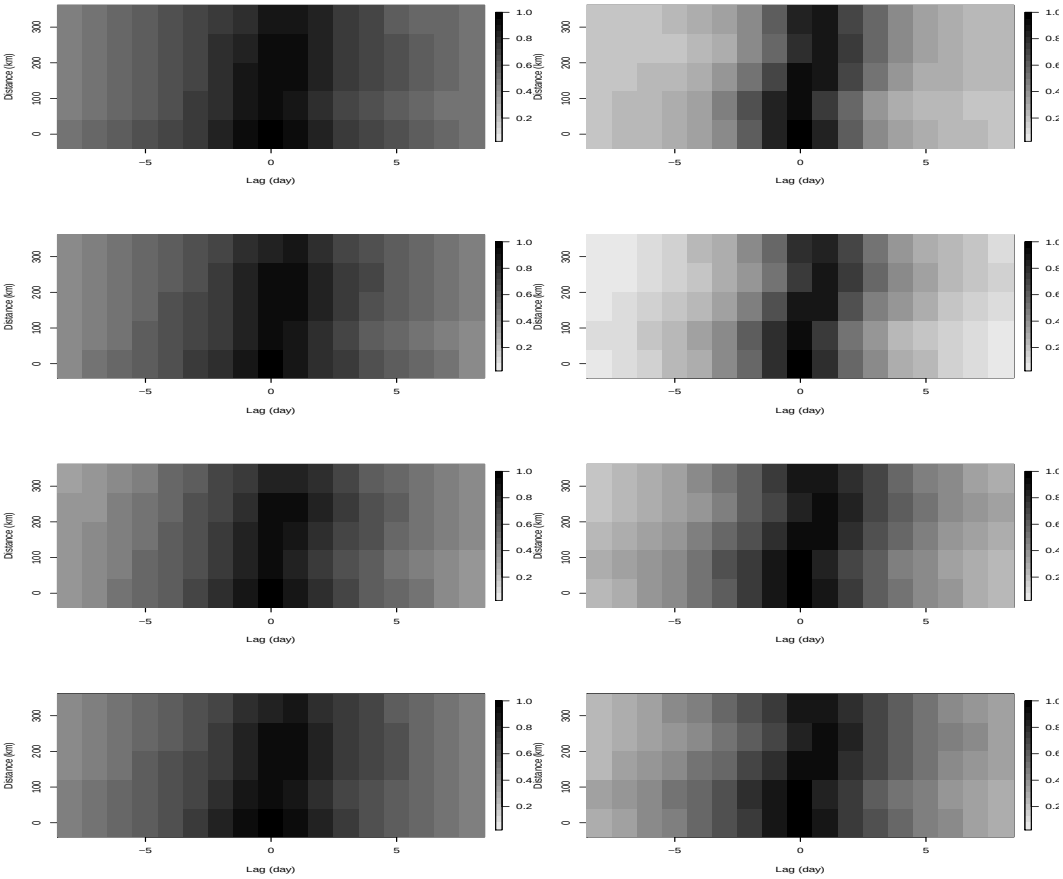

Figure 4.8: Correlation of between $\{u_t\}$ at site 1 and $\{u_t\}$ (left and similar quantities for $\{v_t\}$ on the right) at the other locations at various time-lag. From top panel to the bottom one: data, simulation from VAR(2), AP-MS-VAR$_{C_{Diff(u,v)}}$ and from H-MS-VAR.

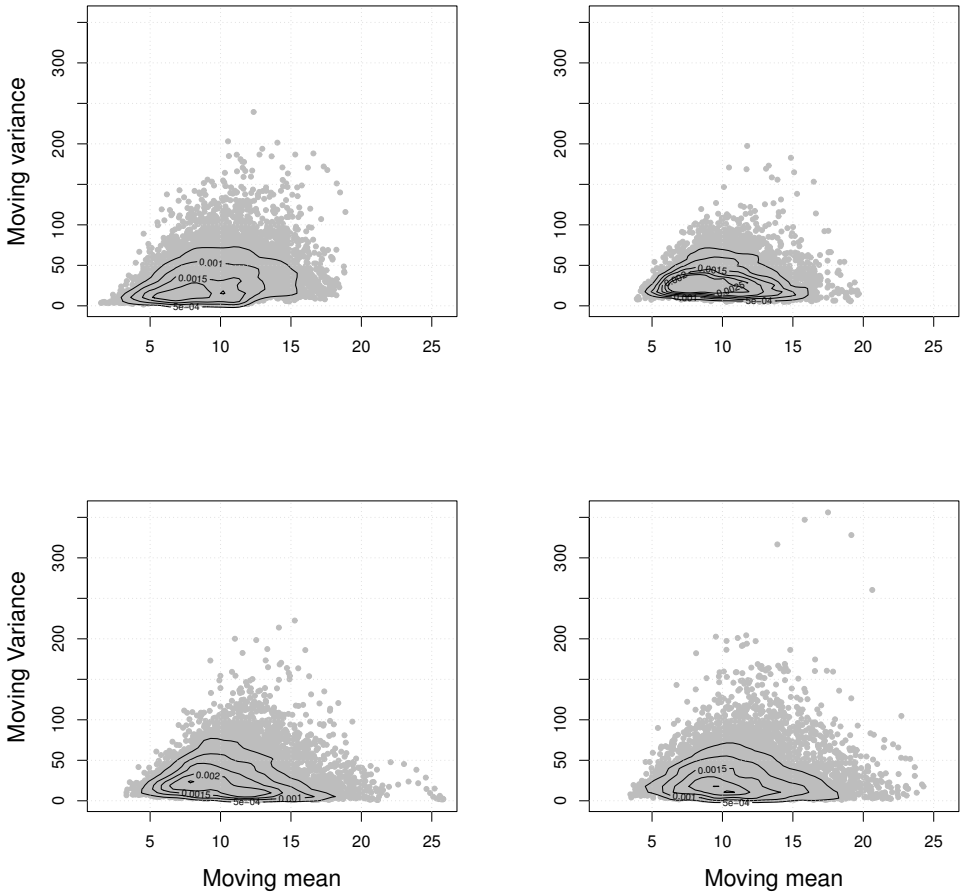

Figure 4.9: Moving variance against of the value $\{U_t\}$ against its moving mean at location 10. From left to right and top to bottom: data, simulation from the VAR(2), AP-MS-VAR$_{C_{Diff(u,v)}}$ and H-MS-VAR

specific regime is acceptable. Nevertheless the construction of these tests is not straightforward since one does not know the distribution of the statistic of test.

We have introduced observed and latent regime-switching framework, we show that both types of regime-switching models have various advantages. On the one hand, 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 we saw in Subsection 4.4.2 where a k-means clustering led to irrelevant regimes. The hidden regime-switching framework seems to overcome this insufficiency by providing regimes that are driven by the conditional distribution and then adapted to the estimation. On the other hand, when considering hidden regime-switching models, the estimation procedure may become challenging when sophisticated models are considered. The extracted regimes are mainly driven by the local data and the proposed conditional distribution, however, they might have less physical interpretation than regimes extracted from atmospheric circulation descriptors. In this work, we show that for the considered dataset, the extracted regimes are meaningful when comparing them to time series of wind and tend to occur in privileged large-scale weather regimes.

Concerning the proposed observed regime-switching models, it seems there is a compromise between meaningful regimes and a good description of the conditional model by a VAR, as highlighted in Section 4.4 when comparing AP-MS-VAR$_{C_{Diff(u,v)}}$ and AP-MS-VAR$_{C_{EOF-(u,v)}}$ models. Indeed we have chosen the AP-MS-VAR$_{C_{Diff(u,v)}}$ since it provides the best BIC index despite its lack of physical interpretation of average fields. This highlights the difficulty to find 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 $C_{Diff(u,v)}$ and similar description of temporal patterns.

When considering other uses than the generation of wind conditions, such as in downscaling, one could improve the modeling of the regimes by using non-homogeneous transitions. Although the choice of the covariate is delicate in the multi-site context. One may use a covariate that is stable across space and time such as the geopotential height $Z500$. Moreover the use of this variable may bring information about inter-annual variability, which is known to be underestimated by most of the weather generators.

Furthermore, finding reduced parametrization of the autoregressive coefficients and of the matrices of covariance of innovations would help to adapt the model to a larger dataset. Indeed when looking at the autoregressive matrices, there are generally privileged predictors according to the regimes which urges the use of constraints matrices in each regime.

Chapter 5

Concluding discussions

We have proposed in this work several stochastic generators of wind conditions off-shore Brittany in France. An originality of these models is that they model two scales, the regional scale through a latent process and the local one. The framework of Markov-switching models and state-space models is used to handle these two scales. In a first time, we have proposed a Gaussian linear state-space model that describes wind speed at several stations. The specificity of this model is to represent a regional wind as a hidden autoregressive process and to downscale this regional wind speed to the local scale through a linear projection. To account for all the information of wind fields, we then consider polar and Cartesian components of wind. We have proposed a single-site model in order to settle properly the modeling of polar and Cartesian components of wind. The modeling is handled into the framework of hidden non-homogeneous Markov-Switching AutoRegressive models. We finally propose a multi-site framework of regime-switching models for Cartesian components of wind at multiple stations. In Section 5.1, we propose a comparison between the proposed multi-site models, in terms of their ability to reproduce some of the observed patterns of wind speed. In Section 5.2, we discuss the proposed work and give some perspectives.

5.1 Comparison of both multi-site models in simulation

The two proposed models involve a latent process which describes the non-observed regional conditions. In the Gaussian linear state-space model (GSSM) of Chapter 2, the latent process is purposed to describe a regional wind that account for space-time motions of wind events. In Chapter 4, the latent process of the H-MS-VAR model is intended to describe the unobserved weather type and it is described by a Markov chain. In Chapter 2, the local scale is written as a linear projection of various lagged versions of the regional process whereas

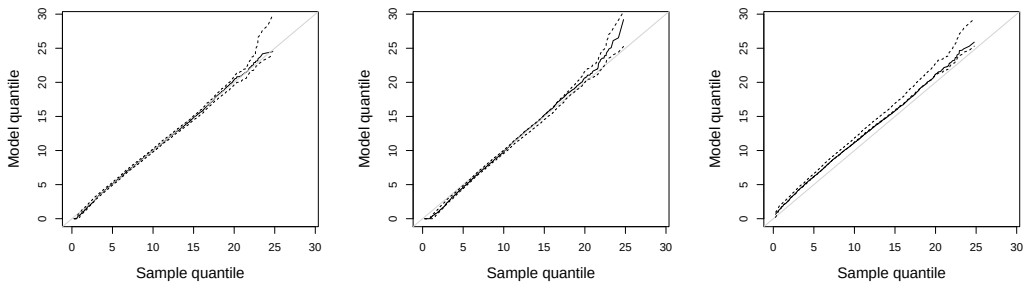

Figure 5.1: Quantile-quantile plot of data and simulated data at the location $(47.25°N, 9.75°W)$. From left to right: simulation from the Gaussian state-space model, from the H-MS-VAR model fitted on $\{\boldsymbol{U}_t\}$ and from the H-MS-VAR model fitted on $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$.

in Chapter 4, the local conditions have their own dynamic conditionally to the regional scale.

The proposed Gaussian linear system of Chapter 2 has been fitted to the reduced dataset of wind speed studied in Chapter 4. The model H-MS-VAR is fitted on $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$ coordinates and wind speed is obtained from the simulation of this process. For comparison purpose, the model H-MS-VAR proposed in Chapter 4 has been fitted on Box-Cox transformed wind speed $\{\boldsymbol{U}_t\}$. We compare here the abilities of these models to reproduce the distribution of wind speed and space-time motions of wind events.

In Figure 5.1, the quantile-quantile plots for the simulated data from the models are depicted. The description of the distribution is very satisfying from the GSSM. The H-MS-VAR model fitted on $\{\boldsymbol{U}_t\}$ is less accurate especially for very small and high values. The one from the H-MS-VAR model for $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$ tends to overestimate all the values of the data. Nevertheless, the distribution of $\{u_t, v_t\}$ is well reproduced at each site by this model. This model is not fitted on the data of $\{\boldsymbol{U}_t\}$ which may deteriorate the description of this distribution.

In Figure 5.2, correlation of wind speed against distance and temporal lags is depicted, we can notice that both models perform almost similarly in reproducing the general shape of the covariance. The non-separability, which is reflected by the asymmetry around the vertical axis at lag 0, is captured. Nevertheless the GSSM and the H-MS-VAR model for $\{\boldsymbol{U}_t\}$ tend to underestimate the temporal dependence around two days of lag. The model for $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$ seems to capture less accurately the spatial structure than the model specific to $\{\boldsymbol{U}_t\}$.

In Figure 5.3, moving mean square errors of wind (with respect to its mov-

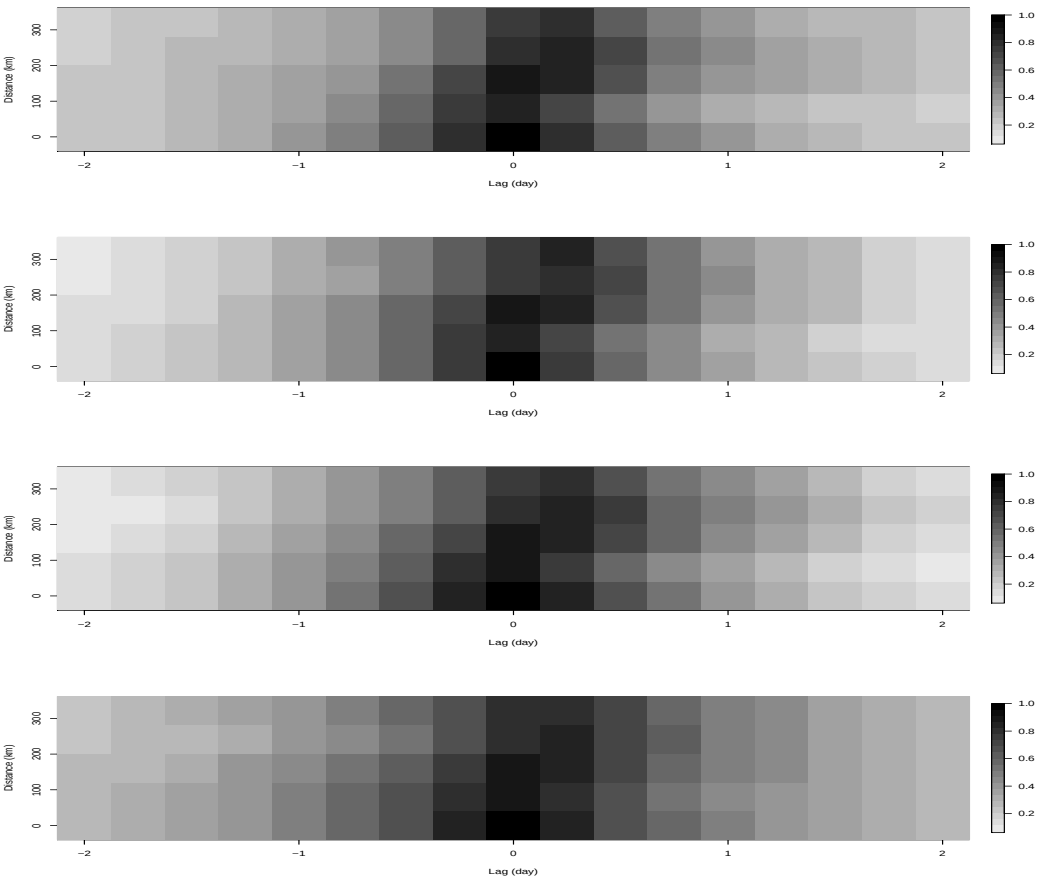

Figure 5.2: Correlation of between $\{\boldsymbol{U}_t\}$ at site 1 and $\{\boldsymbol{U}_t\}$ at the other locations at various time-lag. From top panel to the bottom one: data, simulation from the Gaussian state-space model, from H-MS-VAR fitted on $\{\boldsymbol{U}_t\}$ and from the H-MS-VAR fitted on $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$.

ing average) are computed on interval of time of two days and are depicted. The global shape is better reproduced by the hidden Markov-switching models. This was expected since the H-MS-VAR models discriminate periods of different temporal variability whereas the GSSM model does not enable changes of variances since the model may be written as an ARMA model. However as said previously, the H-MS-VAR model for $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$ has difficulties to capture the range of intensity of $\{\boldsymbol{U}_t\}$, but the tendency to simulate high temporal variability associated to high values of intensity is reproduced. The H-MS-VAR model for $\{\boldsymbol{U}_t\}$ tends to simulate too much variable wind associated with a weak intensity. This is due to the third regime that is associated to a weak mean and a high temporal variability (see Figure 5.4).

In Figure 5.4, moving mean square error of wind with respect to its moving average and moving means are depicted in the associated regime. We compare the ability of the H-MS-VAR models fitted on $\{\boldsymbol{U}_t\}$ and on $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$ to separate the regimes. They seem to be more distinct in terms of intensity and variability of wind speed when the model is fitted on the Cartesian components of wind. The whole information of wind fields, especially wind direction that is a good descriptor of synoptic conditions, enables to extract more meaningful regimes.

Combining a regime-switching framework and a GSSM model instead of a H-MS-VAR model would lighten the modeling of the conditional distribution. Conditionally to the regime, which is driven by a Markov chain, the observations would be described by a GSSM model. Modeling conditional distribution by a GSSM rather than by a VAR model gives a simple and accurate description of the dynamic and is more parsimonious. Additionally parameters of multi-variate AR models are more difficult to interpret than the one of a GSSM model. This modeling would involve huge modeling and computational work due to the estimation procedure that is a current challenging research topic.

5.2 General discussions and perspectives

• *Discussions on statistical and modeling issues.*

In this work, we have introduced models with latent variables of different natures. The first one involves a continuous-valued latent variable. We have studied its identifiability under an original point of view, through the study of the second order structure of the Gaussian process of observations. We have also implemented and compared two methods of estimation of this model. One is based on the method of moments and the other on maximum likelihood. Various reduced models have been proposed and we obtained con-

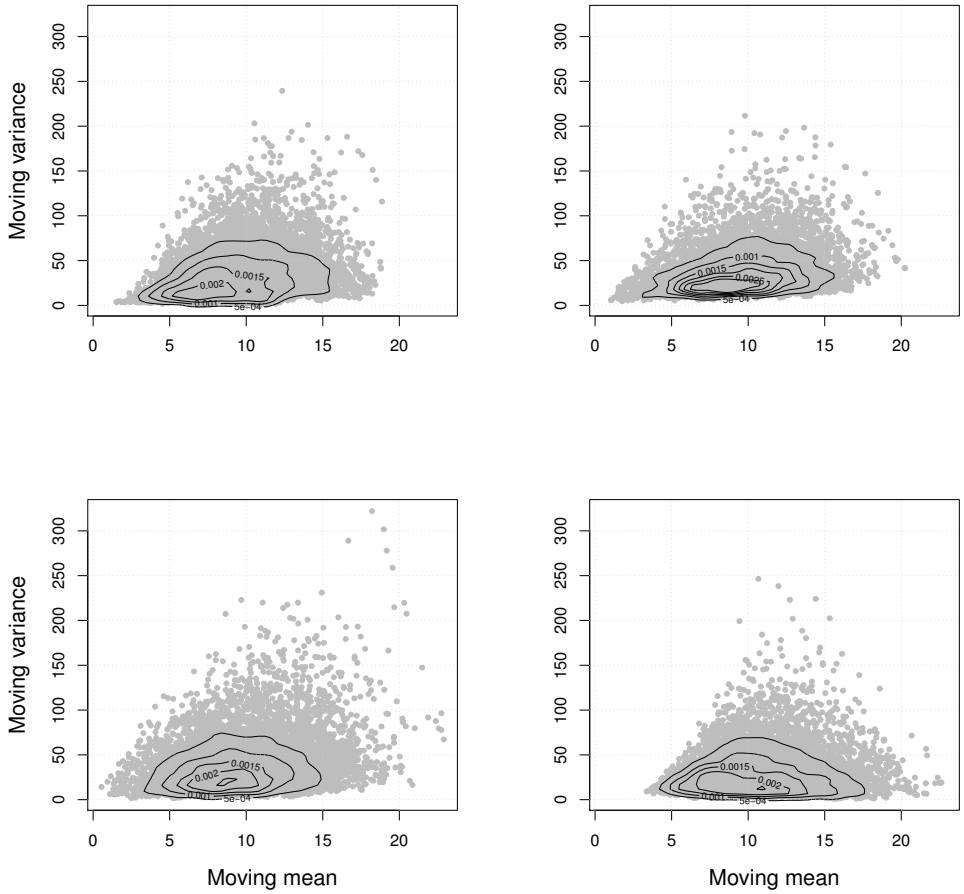

Figure 5.3: Moving mean square errors of wind (with respect to its moving average) against the moving mean of $\{\boldsymbol{U}_t\}$ at the location $(47.25°N, 9.75°W)$. From left top corner to right bottom one: data, simulation from the Gaussian linear system of Chapter 2, from the H-MS-VAR model fitted on $\{\boldsymbol{U}_t\}$ and from the model of Chapter 4.

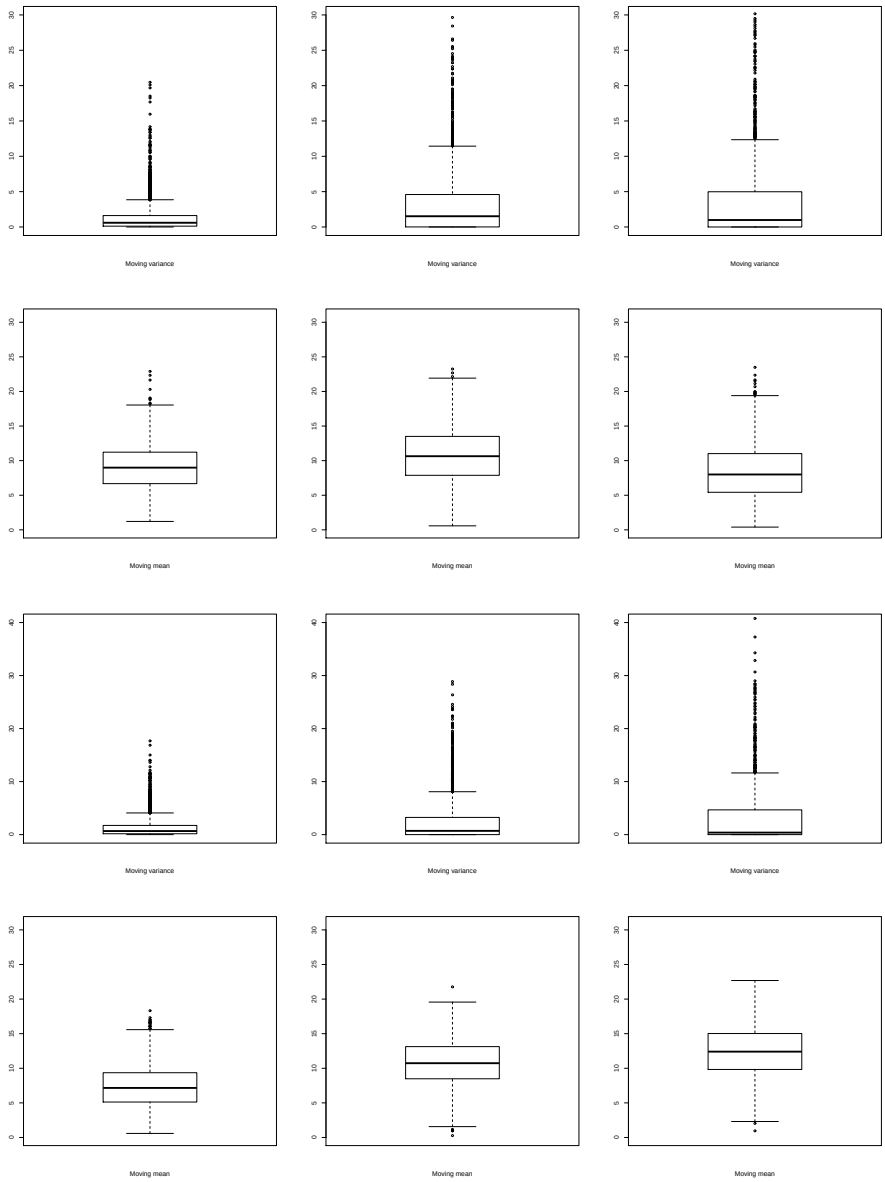

Figure 5.4: Boxplots of moving mean square errors of wind (with respect to its moving average) (first and third lines) and moving means (second and fourth lines) of $\{\boldsymbol{U}_t\}$ in each regime extracted from the H-MS-VAR fitted on $\{\boldsymbol{U}_t\}$ (top lines) and fitted on $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$ (bottom lines).

trasted results. This suggest that some parameters can hardly be described by simple parametric shapes. Further investigations could be led to account for the spatial structure of these parameters. Similarly, in Chapter 4 appropriate parameterization of autoregressive matrices should be investigated to prevent over-parameterization. Besides each regime is associated with privileged predictors which leads to a challenging parameterization.

The proposed Gaussian linear state-space model reveals good abilities to reproduce the average and short-term behaviors of the data through simple dynamic frameworks. However instantaneous temporal variability of wind is not captured by this model. To overcome this insufficiency, we could model the regional wind by a multi-dimensional process, which would be easy to implement. Another possibility would be to add an extra layer modeling the regional weather type as a latent Markov chain. However it is very challenging from the modeling and inference point of view. Indeed the estimation of regime-switching Gaussian linear state-space model is still an on-going research topic.

In Chapter 3, we have introduced a linear-circular model with hidden regime-switching that embeds the dependence between the circular variable and the linear one and the temporal non-linearities of the time series. An originality of this model is that transitions between the shifts are driven by the circular variable, using a proper parameterization. The wind direction is modeled by a hidden Markov-Switching von Mises process with non-homogeneous transitions, and conditionally to this circular variable, wind speed is modeled by a non-homogeneous hidden MS-AR model.

In Chapter 4, Cartesian components of wind are modeled into a multivariate framework of Markov-switching models. Various questions related to regime-switching models have been addressed. First, is discussed the modeling of site-specific regimes against a regional regime. According to qualitative criteria, we conclude that for this dataset the use of a regime common to all locations is reasonable. Some further works would be to build test procedures to investigate the appropriateness of a regional, sub-regional or site-specific regime and give quantitative measures of this appropriateness.

We have proposed for the studied dataset a comparison between the use of observed and latent shifts. To the best of our knowledge, such comparisons have not been conducted before. We proposed several observed regime-switching models, which are based on different classifications. We have discussed the choice of descriptors of the data and of the clustering method in terms of appropriateness to the estimation and simulation of the conditional model in an autoregressive framework. The hidden regime-switching framework seems to provide a good compromise by providing regimes that are adapted to the conditional autoregressive models.

Non-linear behaviors are frequent in time series analysis and especially regime-switchings. To the best of our knowledge, few criteria have been pro-

posed to detect transitions between regimes and quantify the changes in intensity and temporal variability of time series. In that goal, we propose to study the processes of moving means and variances and show that they are good descriptors of the non-linearity of the studied time-series. Through the study of these processes, we highlight the benefit of using regime-switching in the modeling of the alternate of intensity and temporal variability in wind conditions.

• *Discussions about the simulation results.*

In terms of applications, we have proposed a single-site and several multi-site models for wind conditions. Very few multi-site stochastic generators of wind conditions have been constructed in the literature and especially for the zonal and meridional coordinates of wind.

The distribution of wind speed is well described by the models proposed in Chapters 2 and 3. In Chapter 3, the distribution of wind direction and of the Cartesian components of wind are well reproduced by the proposed single-site models. In this chapter, the use of non-homogeneous transitions between regimes enables to describe more accurately the various distributions than when homogeneous transitions are considered. We have proposed a single-site framework in Chapter 3 that encompasses a great part of wind speed and direction dependences. In the multi-site context of Chapter 4, the distribution of Cartesian components of wind is still well described by the proposed models.

The multi-site models, proposed in Chapters 2 and 4, reveal good abilities to reproduce the space-time covariance structure of the considered processes. Namely the patterns of non-separability and non-spatial stationarity are well accounted by the different models. Moreover, the marginal temporal dependence of the various processes is also captured by the associated models. Especially we highlight the benefit of non-homogeneous transitions in Chapter 3 and of accounting for all the information of wind fields in Subsection 5.1.

In order to model the alternate of different intensity and temporal variability in wind conditions, we have proposed in Chapter 4 several *a priori* regime-switching models and a hidden one and we compare them. The observed regime-switching models are based on several classifications extracted from a large-scale descriptor of atmospheric circulation and from the local wind conditions. We discuss the difficulty to find physically consistent *a priori* regimes that are also appropriate to the description of the conditional model in an autoregressive framework. The hidden regime-switching framework seems the most appropriate to respond this compromise by providing interpretable regimes and an accurate description in simulation. Finally we highlight the benefit of using regime-switching models in the description of the alternate of different intensity and temporal variability in wind conditions. We show that the hidden Markov-switching model is the most able to reproduce these features among all the proposed regime-switching models. In Subsection 5.1, we

have seen that the Gaussian linear state-space model proposed in Chapter 2 is not able to reproduce well these instantaneous behaviors, due to the proposed modeling by a Gaussian process with a constant variance.

Involving a regime-switching framework enables to capture more information than involving a latent continuous state like in Chapter 2's framework. Indeed it allows to capture more than the average space-time motions, it enables to extract periods with typical patterns of intensity, temporal variability and dependence between intensity and direction. In Subsection 5.1, weather states seem to be better described when they are extracted from a model that involves direction and intensity of wind. However the model fitted on $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$ is less accurate in the modeling of wind speed, this is observed with the single-site and multi-site models. One difficulty is to find a model with distinct and relevant regimes that captures well several margins like the intensity of wind. One may use adequate estimation procedures to fit the H-MS-VAR model on $\{\boldsymbol{u}_t, \boldsymbol{v}_t\}$ with appropriate constraints that improve the modeling of intensity of wind. Indeed the maximum likelihood method focuses on the distribution $p(\boldsymbol{Y}_t|\boldsymbol{Y}_{t-1})$, one may add constraints to reproduce some parameters associated with the marginal distribution of $\boldsymbol{Y}$.

For the studied dataset and the proposed models, short-term prediction have been used as a validation. The proposed models are designed for generating artificial sequences of wind, then forecast is not their main features. Besides beyond six hours ahead, statistical models are not recommended for wind prediction, models based on physics are rather preferred (Giebel et al., 2011). Nevertheless, the proposed models behave in short-term prediction more accurately than the benchmark persistence forecast, especially due to the accounting of spatial structure. Prediction of wind speed has known a wide development those last decades, the literature is wide on the area (see (Costa et al., 2008; Giebel et al., 2011) for reviews on models for wind short-term prediction). Similarly to stochastic generators, recent statistical methodologies for wind prediction turned toward space-time frameworks and non-linear models. Space-time methodologies have also been developed to analyze wind power or wind prediction data (Tastu et al., 2011; Girard and Allard, 2012) and non-linear models are proposed for instance in (Zhu and Genton, 2012; Pinson et al., 2008).

To conclude, the literature is rich on single-site models for wind and these generators work quite accurately for the proposed temporal scale, like the model proposed in Chapter 3. Some further works should concentrate on the modeling of finer time scale like hourly or sub-hourly scales. Few generators have been proposed to model very fine time scales. For instance, the wind energy field requires very fine time scales (Giebel et al., 2011). The modeling has to be adapted when considering a finer time scale since the correlation between observations is stronger and interactions between variables might be accounted in a different way.

A lot of research work remains on multi-site wind models. In particular, parsimonious models that capture the complex spatial-temporal information of wind and its regime-switching patterns are needed. This field is still an open research area. Besides accounting for and modeling other meteorological variables would help to get information from the synoptic conditions but the modeling of interactions between these variables at several locations is very challenging.

Bibliography

Abrahamsen, P. (1997). *A review of Gaussian random fields and correlation functions*. Norsk Regnesentral/Norwegian Computing Center.

Ailliot, P., Besssac, J., Monbet, V., and Pène, F. (2014). Non-homogeneous hidden markov-switching models for wind time series.

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

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

Ailliot, P., Monbet, V., and Prevosto, M. (2006b). An autoregressive model with time-varying coefficients for wind fields. *Environmetrics*, 17(2):107–117.

Ailliot, P. and Pène, F. (2013). Consistency of the maximum likelihood estimate for non-homogeneous markov-switching models.

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

Bai, J. and Wang, P. (2012). Identification and estimation of dynamic factor models.

Bardossy, A. and Plate, E. J. (1992). Space-time model for daily rainfall using atmospheric circulation patterns. *Water Resources Research*, 28(5):1247–1259.

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

Bork, L. (2010). *Macro factors, Monetary policy analysis and affine term structure models.* Aarhus School of Business, Department of Business Studies.

Breckling, J. (1989). *The analysis of directional time series: applications to wind speed and direction.* Lecture Notes in Statistics. Springer-Verlag Berlin.

Brockwell, P. J. and Davis, R. A. (2002). *Introduction to time series and forecasting, second edition.* Springer-Verlag, New York.

Brockwell, P. J. and Davis, R. A. (2006). *Time series: theory and methods.* Springer Series in Statistics. Springer, New York. Reprint of the second (1991) edition.

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

Caines, P. E. (1988). *Linear stochastic systems.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York.

Cappé, O., Moulines, E., and T., R. (2005). *Inference in hidden Markov models.* Springer-Verlag, New York.

Cassou, C. (2008). Intraseasonal interaction between the madden–julian oscillation and the north atlantic oscillation. *Nature*, 455(7212):523–527.

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

Cattiaux, J., Douville, H., and Peings, Y. (2013). European temperatures in cmip5: origins of present-day biases and future uncertainties. *Climate Dynamics*, 41(11-12):2889–2907.

Costa, A., Crespo, A., Navarro, J., Lizcano, G., Madsen, H., and Feitosa, E. (2008). A review on the young history of the wind power short-term prediction. *Renewable and Sustainable Energy Reviews*, 12(6):1725–1744.

Cox, D. R. and Isham, V. (1988). A simple spatial-temporal model of rainfall. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 415(1849):317–328.

Cressie, N. A. C. (1991). *Statistics for spatial data.* Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York. A Wiley-Interscience Publication.

de Luna, X. and Genton, M. G. (2005). Predictive spatio-temporal models for spatially sparse environmental data. *Statist. Sinica*, 15(2):547–568.

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

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

Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods*, volume 38 of *Oxford Statistical Science Series*. Oxford University Press, Oxford, second edition.

Finkenstädt, B., Held, L., and Isham, V., editors (2007). *Statistical methods for spatio-temporal systems*, volume 107 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton, FL. Papers from the 6th Séminaire Européen de Statistique held in Bernried, December 12–18, 2004.

Fisher, N. I. and Lee, A. J. (1983). A correlation coefficient for circular data. *Biometrika*, 70:327–332.

Fisher, N. I. and Lee, J. (1994). Time series analysis of circular data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 56:327–339.

Flecher, C., Naveau, P., Allard, D., and Brisson, N. (2010). A stochastic daily weather generator for skewed data. *Water Resources Research*, 46(7):W07519.

Francq, C. and Roussignol, M. (1998). Ergodicity of autoregressive processes with markov-switching and consistency of the maximum-likelihood estimator. *Statistics*, 32(2):151–173.

Fraser, M. D., Hsu, Y. S., and J., W. J. (1981). Identifiability of finite mixtures of von mises distributions. *Annals of statistics*, 9:1130–1131.

Fuentes, M., Chen, L., Davis, J. M., and Lackmann, G. M. (2005). Modeling and predicting complex space–time structures and patterns of coastal wind fields. *Environmetrics*, 16(5):449–464.

Gabriel, K. R. and Neumann, J. (1962). A markov chain model for daily rainfall occurrence at tel aviv. *Quarterly Journal of the Royal Meteorological Society*, 88(375):90–95.

Giebel, G., Brownsword, R., Kariniotakis, G., Denhard, M., and Draxl, C. (2011). The state-of-the-art in short-term prediction of wind power: A literature overview. Technical report, ANEMOS. plus.

Girard, R. and Allard, D. (2012). Spatio-temporal propagation of wind power prediction errors. *Wind Energy*.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space-time data. *J. Amer. Statist. Assoc.*, 97(458):590–600.

Gneiting, T., Larson, K., Westrick, K., Genton, M. G., and Aldrich, E. (2006). Calibrated probabilistic forecasting at the stateline wind energy center: The regime-switching space–time method. *Journal of the American Statistical Association*, 101(475):968–979.

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

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

Hannan, E. and Deistler, M. (1988). *The statistical theory of linear systems*. Springer Texts in Statistics. John Wiley, New York, second edition. With 1 CD-ROM (Windows).

Haskard, K. A. (2007). An anisotropic matérn spatial covariance model: Reml estimation and properties. *Ph.D. dissertation, University of Adelaide, Australia*.

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

Hering, A. S. and Genton, M. G. (2010). Powering up with space-time wind forecasting. *Journal of the American Statistical Association*, 105(489):92–104.

Hinkley, D. (1977). On quick choice of power transformation. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 26(1):pp. 67–69.

Hofmann, M. and Sperstad, I. B. (2013). Nowicob–a tool for reducing the maintenance costs of offshore wind farms. *Energy Procedia*, 35:177–186.

Holzmann, H., Munk, A., Suster, M., and Zucchini, W. (2006). Hidden markov models for circular and linear-circular time series. *Environmental and Ecological Statistics*, 13(3):325–347.

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

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

Kamal, L. and Jafri, Y. Z. (1997). Time series models to simulate and forecast hourly averaged wind speed in quetta, pakistan. *Solar Energy*, 61(1):23–32.

Kato, S. (2010). A markov process for circular data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72:655–672.

Khalili, M., Leconte, R., and Brissette, F. (2007). Stochastic multisite generation of daily precipitation data using spatial autocorrelation. *Journal of Hydrometeorology*, 8(3):396–412.

Kleiber, W., Katz, R. W., and Rajagopalan, B. (2012). Daily spatiotemporal precipitation simulation using latent and transformed gaussian processes. *Water Resources Research*, 48(1):n/a–n/a.

Kleiber, W., Katz, R. W., and Rajagopalan, B. (2013). Daily minimum and maximum temperature simulation over complex terrain. *The Annals of Applied Statistics*, 7(1):588–612.

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

Lei, M., Shiyan, L., Chuanwen, J., Hongling, L., and Yan, Z. (2009). A review on the forecasting of wind speed and generated power. *Renewable and Sustainable Energy Reviews*, 13(4):915–920.

Ljung, L. (1999). *System identifiability*. Springer Texts in Statistics. Prentice Hall, New Jersey, second edition. With 1 CD-ROM (Windows).

Maraun, D., Wetterhall, F., Ireson, A., Chandler, R., Kendon, E., Widmann, M., Brienen, S., Rust, H., Sauter, T., Themeßl, M., et al. (2010). Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, 48(3).

Mardia, K. V. (1972). *Statistics of directional data*. Academic press, New York.

McDonald, I. L. and Zucchini, W. (1997). *Hidden Markov and other models for discrete-valued time series*. Chapman & Hall/CRC, London.

Michelangeli, P. A., Vautard, R., and Legras, B. (1995). Weather regimes: recurrence and quasi stationarity. *Journal of the Atmospheric Sciences*, 52(8):1237–1256.

Milliff, R. F., Bonazzi, A., Wikle, C. K., Pinardi, N., and Berliner, L. M. (2011). Ocean ensemble forecasting. part i: Ensemble mediterranean winds from a bayesian hierarchical model. *Quarterly Journal of the Royal Meteorological Society*, 137(657):858–878.

Modlin, D., Fuentes, M., and Reich, B. (2012). Circular conditional autoregressive modeling of vector fields. *Environmetrics*, 23(1):46–53.

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

Muñoz, M. P., Sànchez-Espigares, J. A., and D., M. M. (2013). Non linear statistical models to improve wind power forecasts. *Technical report*.

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

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of econometrics, Vol. IV*, volume 2 of *Handbooks in Econom.*, pages 2111–2245. North-Holland, Amsterdam.

Nfaoui, H., Buret, J., and Sayigh, A. A. M. (1996). Stochastic simulation of hourly average wind speed sequences in tangiers (morocco). *Solar Energy*, 56(3):301–314.

Papadopoulos, P. and Digalakis, V. (2010). Identification of linear systems in canonical form through an em framework. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 4110–4113. IEEE.

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

Qin, X., Zhang, J., and Yan, X. (2010). A new circular distribution and its application to wind data. *Journal of Mathematics Research*, 2(3).

Racsko, P., Szeidl, L., and Semenov, M. (1991). A serial approach to local stochastic weather models. *Ecological modelling*, 57(1):27–41.

Rajagopalan, B. and Lall, U. (1999). A k-nearest-neighbor simulator for daily precipitation and other weather variables. *Water resources research*, 35(10):3089–3101.

Refice, A., Belmonte, A., Bovenga, F., and Pasquariello, G. (2011). On the use of anisotropic covariance models in estimating atmospheric dinsar contributions. *Geoscience and Remote Sensing Letters, IEEE*, 8(2):341–345.

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

Rychlik, I. and Mustedanagic, A. (2013). A spatial-temporal model for wind speeds variability.

Šaltytė Benth, J. and Šaltytė, L. (2011). Spatial-temporal model for wind speed in Lithuania. *J. Appl. Stat.*, 38(6):1151–1168.

Semenov, M. A. and Barrow, E. M. (1997). Use of a stochastic weather generator in the development of climate change scenarios. *Climatic change*, 35(4):397–414.

Shumway, R. H. and Stoffer, D. S. (2006). *Time series analysis and its applications*. Springer Texts in Statistics. Springer, New York, second edition. With R examples.

Skidmore, E. and Tatarko, J. (1990). Stochastic wind simulation for erosion modeling. *Transactions of the ASAE*, 33(6):1893–1899.

Srikanthan, R. and McMahon, T. (1999). Stochastic generation of annual, monthly and daily climate data: A review. *Hydrology and Earth System Sciences*, 5(4):653–670.

Tastu, J., Pinson, P., Kotwa, E., Madsen, H., and Nielsen, H. A. (2011). Spatio-temporal analysis and modeling of short-term wind power forecast errors. *Wind Energy*, 14(1):43–60.

Tastu, J., Pinson, P., and Madsen, H. (2013). Space-time scenarios of wind power generation produced using a gaussian copula with parametrized precision matrix. Technical report, Technical University of Denmark.

Thompson, C. S., Thomson, P. J., and Zheng, X. (2007). Fitting a multisite daily rainfall model to new zealand data. *Journal of Hydrology*, 340(1):25–39.

Tong, H. (1990). *Non-linear time series: a dynamical system approach*. Oxford University Press, Oxford, U.K.

Vautard, R. (1990). Multiple weather regimes over the north atlantic: Analysis of precursors and successors. *Monthly Weather Review*, 118(10):2056–2081.

Vrac, M., Stein, M., and Hayhoe, K. (2007). Statistical downscaling of precipitation through nonhomogeneous stochastic weather typing. *Climate Research*, 34(3):169.

Wikle, C. K. and Hooten, M. B. (2010). A general science-based framework for dynamical spatio-temporal models. *Test*, 19(3):417–451.

Wikle, C. K., Milliff, R. F., Nychka, D., and Berliner, L. M. (2001). Spatiotemporal hierarchical bayesian modeling tropical ocean surface winds. *Journal of the American Statistical Association*, 96(454):382–397.

Wilks, D. S. (1992). Adapting stochastic weather generation algorithms for climate change studies. *Climatic Change*, 22(1):67–84.

Wilks, D. S. (1998). Multisite generalization of a daily stochastic precipitation generation model. *Journal of Hydrology*, 210(1):178–191.

Wilks, D. S. (1999). Simultaneous stochastic simulation of daily precipitation, temperature and solar radiation at multiple sites in complex terrain. *Agricultural and Forest Meteorology*, 96(1):85–101.

Wilks, D. S. (2009). A gridded multisite weather generator and synchronization to observed weather data. *Water Resources Research*, 45(10).

Wilks, D. S. (2010). Use of stochastic weather generators for precipitation downscaling. *Wiley Interdisciplinary Reviews: Climate Change*, 1(6):898–907.

Wilks, D. S. (2012). Stochastic weather generators for climate-change downscaling, part ii: multivariable and spatially coherent multisite downscaling. *Wiley Interdisciplinary Reviews: Climate Change*, 3(3):267–278.

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

Wilson, L. L., Lettenmaier, D. P., and Skyllingstad, E. (1992). A hierarchical stochastic model of large-scale atmospheric circulation patterns and multiple station daily precipitation. *Journal of Geophysical Research: Atmospheres (1984–2012)*, 97(D3):2791–2809.

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

Yang, C., Chandler, R. E., Isham, V. S., and Wheater, H. S. (2005). Spatial-temporal rainfall simulation using generalized linear models. *Water Resources Research*, 41(11).

Zhu, X. and Genton, M. G. (2012). Short-term wind speed forecasting for power system operations. *International Statistical Review*, 80(1):2–23.

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

Zucchini, W. and MacDonald, I. (2009). *Hidden Markov Models for time series: an introduction using R*. Number 110 in Monographs on statistics and applied probability. CRC Press.