Gaussian linear state-space model for wind fields in the North-East Atlantic

Julie Bessac, Pierre Ailliot, Valérie Monbet

Gaussian linear state-space model for wind fields in the North-East Atlantic

J. Bessac[1], P. Ailliot[2], V. Monbet [1,3]

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

[2] *Laboratoire de Mathématiques de Bretagne Atlantique, UMR 6205, Université de Brest, Brest, France*

[3] *INRIA Rennes, ASPI, Rennes, France*

Correspondence to: Ms J. Bessac, Institut de Recherche Mathématiques de Rennes, UMR 6625, Université de Rennes 1, Rennes, 35000, France, julie.bessac at univ-rennes1.fr

Article category: research

Running head: Gaussian linear state-space model for wind fields

Abstract

A multisite 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 moments and the EM algorithm. 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.

Keywords: Multisite wind generators, Space-time model, State-space model, EM algorithm, Identifiability.

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 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 paper, a structural model that aims at simulating wind speed at several locations is investigated. The main idea consists in introducing a latent variable that aims at describing regional wind conditions and the observed local wind is modeled as a function of the regional wind at different lags in order to reproduce the mean displacement of the air masses. The framework is kept simple with a linear Gaussian model 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 paper are presented in Section 2. The model is described in Section 3. Parameter estimation and fitting procedures are also discussed in this section. The model is validated in Section 4 and it is shown in particular 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 5. Conclusions are given in Section 6. Parameter identifiability and non full-symmetry

are proven in the supplementary materials.

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 observations with numerical weather prediction models, provides a relevant alternative for meteorological or climatological studies. In this paper 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 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 paper could easily be adapted to handle datasets with a 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 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 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 moves are generally eastward. It induces non-separability and non full-symmetry properties of the space-

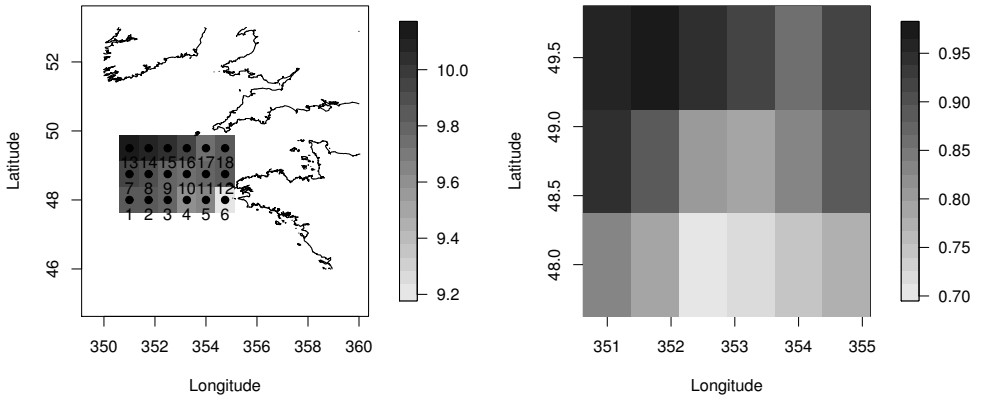

Figure 1: Left panel: mean wind speed at the 18 numbered points under study in the North-East Atlantic. Right panel: estimated values of the power in the Box-Cox method at the 18 locations.

time covariance function of the wind speed as for the dataset of wind speed in Ireland considered in (Haslett and Raftery, 1989; Gneiting, 2002). The lagged by 1 cross-correlations shown in Figure 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'; in average western locations see the meteorological events before the eastern locations. This asymmetry is less pronounced in latitude but reveals flows from north to south. Furthermore, the correlations reveal some anisotropy as dependences in latitude and longitude differ (see Figure 2).

Wind speed distribution is known to be skewed. It is often modeled as a Weibull distribution (Brown et al., 1984) but other distributions such as the skew normal distribution have also been considered (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

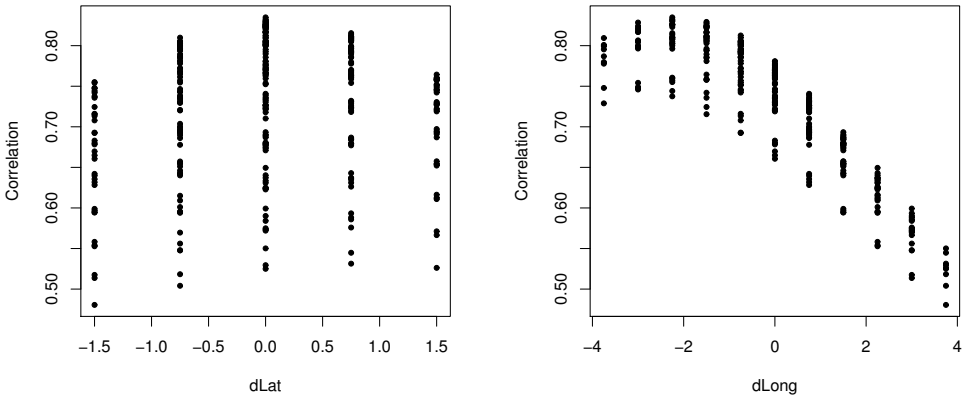

Figure 2: Lagged-one cross-correlations against differences of latitude (left) and longitude (right).

approximately Gaussian marginal distribution. This method has been extensively used for analyzing wind time series at a single location, see for example (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

$$S(\lambda_i) = \frac{\text{mean}(y_{\lambda_i,i,t}) - \text{median}(y_{\lambda_i,i,t})}{\sqrt{\text{var}(y_{\lambda_i,i,t})}}. \tag{1}$$

The resulting estimates are shown in Figure 1 with values ranging from about 1 (Gaussian distribution) in the north-west to 0.7 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 as done in (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 1. The simulation results given in Section 4 (see Figure 5) indicate that this simple transformation permits to reproduce the marginal distributions of the wind data considered in this study.

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 (Durbin and Koopman, 2012; Brockwell and Davis, 2006) and space-time processes (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.

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 (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 prevail-

ing 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 paper, is defined as

$$(\mathrm{M}) \begin{cases} X_{t+1} &= \rho X_t + \sigma \epsilon_{t+1}, \\ \boldsymbol{Y}_t &= \boldsymbol{\alpha}_1 X_{t+1} + \boldsymbol{\alpha}_0 X_t + \boldsymbol{\alpha}_{-1} X_{t-1} + \boldsymbol{\Gamma}^{1/2} \boldsymbol{\eta}_t \end{cases} \quad \text{for } t \geq 0,$$

$\boldsymbol{Y}_t \in \mathbb{R}^K$ is the observed process, its K coordinates correspond to the mean-corrected transformed wind speed at the $K = 18$ locations. $\{\epsilon_t\}$ and $\{\boldsymbol{\eta}_t\}$ are independent Gaussian white noises with zero-means and identity covariance matrices. $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_{-1}$ are K-dimensional vectors that link the lagged values of the regional process $\{X_t\}$ to local wind conditions. The covariance matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{K \times K}$ models the spatial structure of small-scale fluctuations. In finance and economics or when high dimensional data are considered this covariance matrix is often assumed to be diagonal (Wikle and Hooten, 2010). 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 5.1. In the sequel, we denote $\boldsymbol{\Lambda} = (\boldsymbol{\alpha}_1 | \boldsymbol{\alpha}_0 | \boldsymbol{\alpha}_{-1}) \in \mathbb{R}^{K \times 3}$ and $\theta = (\rho, \sigma, \boldsymbol{\Lambda}, \boldsymbol{\Gamma})$ 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

$\{\boldsymbol{Y_t}\}$.

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 non-identifiable without additional constraints (Hannan and Deistler, 1988; Ljung, 1999; Bai and Wang, 2012; Bork, 2010). Identifiability of linear Gaussian state-space models was initially investigated in control theory and has been largely explored during the last decades. However we could not find any result that applies directly to the model considered in this paper.

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

$$\begin{aligned}
\mathrm{cov}_\theta(\boldsymbol{Y}_t, \boldsymbol{Y}_t) &= \frac{\sigma^2}{1-\rho^2}\Big(\boldsymbol{\alpha}_1(\boldsymbol{\alpha}_1 + \rho\boldsymbol{\alpha}_0 + \rho^2\boldsymbol{\alpha}_{-1})^t + \boldsymbol{\alpha}_0(\rho\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_0 + \rho\boldsymbol{\alpha}_{-1})^t + \\
&\quad \boldsymbol{\alpha}_{-1}(\rho^2\boldsymbol{\alpha}_1 + \rho\boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_{-1})^t\Big) + \boldsymbol{\Gamma}, \qquad (2)\\
\mathrm{cov}_\theta(\boldsymbol{Y}_t, \boldsymbol{Y}_{t+1}) &= \frac{\sigma^2}{1-\rho^2}\Big(\boldsymbol{\alpha}_1(\rho\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_0 + \rho\boldsymbol{\alpha}_{-1})^t + \boldsymbol{\alpha}_0(\rho^2\boldsymbol{\alpha}_1 + \rho\boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_{-1})^t + \\
&\quad \rho\boldsymbol{\alpha}_{-1}(\rho^2\boldsymbol{\alpha}_1 + \rho\boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_{-1})^t\Big), \qquad (3)\\
\mathrm{cov}_\theta(\boldsymbol{Y}_t, \boldsymbol{Y}_{t+k}) &= \frac{\sigma^2}{1-\rho^2}\rho^{k-2}(\boldsymbol{\alpha}_1 + \rho\boldsymbol{\alpha}_0 + \rho^2\boldsymbol{\alpha}_{-1})(\rho^2\boldsymbol{\alpha}_1 + \rho\boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_{-1})^t, \qquad (4)\\
&\quad \text{for all } k \geq 2.
\end{aligned}$$

The study of this space-time covariance function leads to the following Proposition, which is proven in the supplementary materials.

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 $\boldsymbol{Y}_t$ 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 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 (2-4) is neither full-symmetric nor separable (see the supplementary materials). 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.

3.3 Parameter estimation

Two methods of estimation have been implemented and compared. The first one is a method of moments based on the second-order structure of the process $\{\boldsymbol{Y_t}\}$ given by (2-4). It consists in numerically minimizing the following

objective function

$$\theta \rightarrow \|\widehat{\text{cov}}(\boldsymbol{Y}_t, \boldsymbol{Y}_t) - \text{cov}_\theta(\boldsymbol{Y}_t, \boldsymbol{Y}_t)\|_2^2 + \|\widehat{\text{cov}}(\boldsymbol{Y}_t, \boldsymbol{Y}_{t+1}) - \text{cov}_\theta(\boldsymbol{Y}_t, \boldsymbol{Y}_{t+1})\|_2^2 \quad (5)$$
$$+\|\widehat{\text{cov}}(\boldsymbol{Y}_t, \boldsymbol{Y}_{t+2}) - \text{cov}_\theta(\boldsymbol{Y}_t, \boldsymbol{Y}_{t+2})\|_2^2 + \|\widehat{\text{cov}}(\boldsymbol{Y}_t, \boldsymbol{Y}_{t+3}) - \text{cov}_\theta(\boldsymbol{Y}_t, \boldsymbol{Y}_{t+3})\|_2^2,$$

where $\widehat{\text{cov}}$ denotes the empirical covariance function and $\|.\|_2$ stands for the matrix Frobenius norm. This method, denoted by GMM for Generalized Method of Moments in the sequel, is usual in geostatistics (Cressie, 1991). We have chosen to consider only the first four lags of the autocovariance function when building the objective function (5). It corresponds to the minimal number of terms needed to identify the parameters (see the supplementary materials). 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). The EM algorithm aims at maximizing the incomplete log-likelihood function

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

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 values, which need to be

chosen carefully. We used the following procedure that 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

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

  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 (3) and (4).

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

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

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

3.4 Properties of the estimates

Under suitable conditions, GMM (Newey and McFadden, 1994) and ML (Newey and McFadden, 1994; Shumway and Stoffer, 2006; Hannan and Deistler, 1988; Caines, 1988) estimates are consistent and asymptotically Gaussian. In order to assess the performances of the estimates for the practical application considered in this paper, 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 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.

4 Results

In order to validate the proposed model we check its physical realism and its ability to generate artificial wind conditions with statistical properties similar to the ones of the dataset. We compare the GMM and ML estimates through

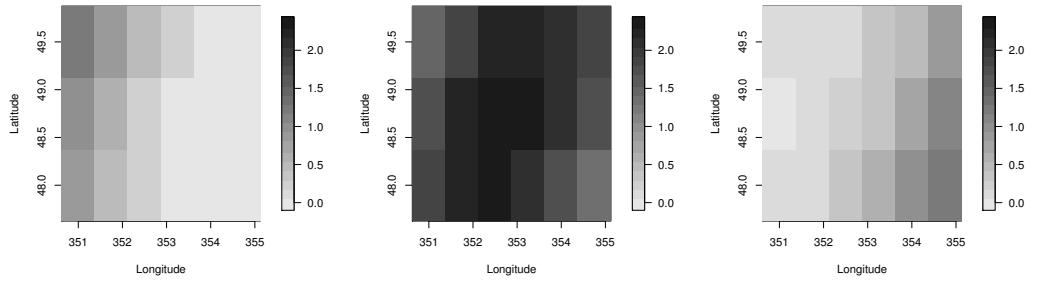

Figure 3: ML estimates of $\boldsymbol{\alpha}_1$ (left panel) $\boldsymbol{\alpha}_0$ (middle panel) and $\boldsymbol{\alpha}_{-1}$ (right panel).

this validation in order to investigate their robustness in a practical context.

4.1 Interpretability

The loading matrix $\boldsymbol{\Lambda}$ links the latent process to the observed wind conditions. The values of $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_{-1}$ shown on Figure 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 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 4). The shape of $\boldsymbol{\Gamma}$ has a block structure that is induced by the geometry of the domain and the numbering of the sites (see Figure 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

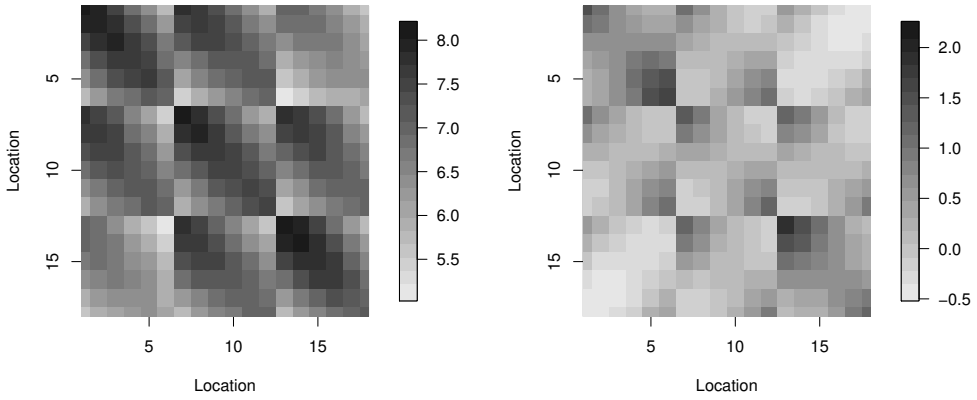

Figure 4: Empirical covariance matrix of the wind data (left) and ML estimates of $\mathbf{\Gamma}$ (right).

at the locations that 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 $\mathbf{\Gamma}$ (see Section 5.1).

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 the quantile-quantile plots shown on Figure 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 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 is designed to make the first lags

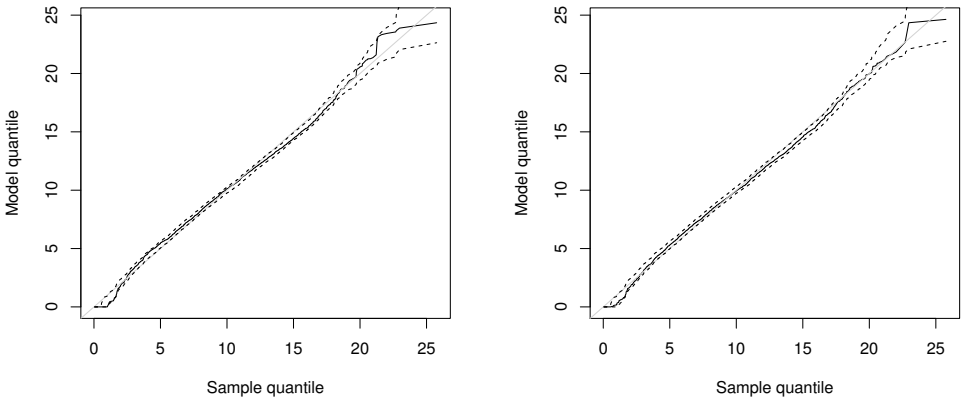

Figure 5: Quantile-Quantile plot at location 9 for the model (M) and the parameters estimated by GMM (left) and ML (right). The dashed lines correspond to 90% prediction intervals computed by simulation.

of the empirical autocovariance function coincide with the one of the fitted model. Figure 6 shows however that the fit is better for lags greater than one day with the ML estimates, which take into account longer term dynamics and leads to a higher value of ρ (0.76 for ML against 0.70 for GMM). The better fit of the ML estimates is also coherent with Table 1. Note also that the models reproduce the time shift between locations 13 and 18 that is induced by the prevailing westerly flow (see Figure 6).

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.

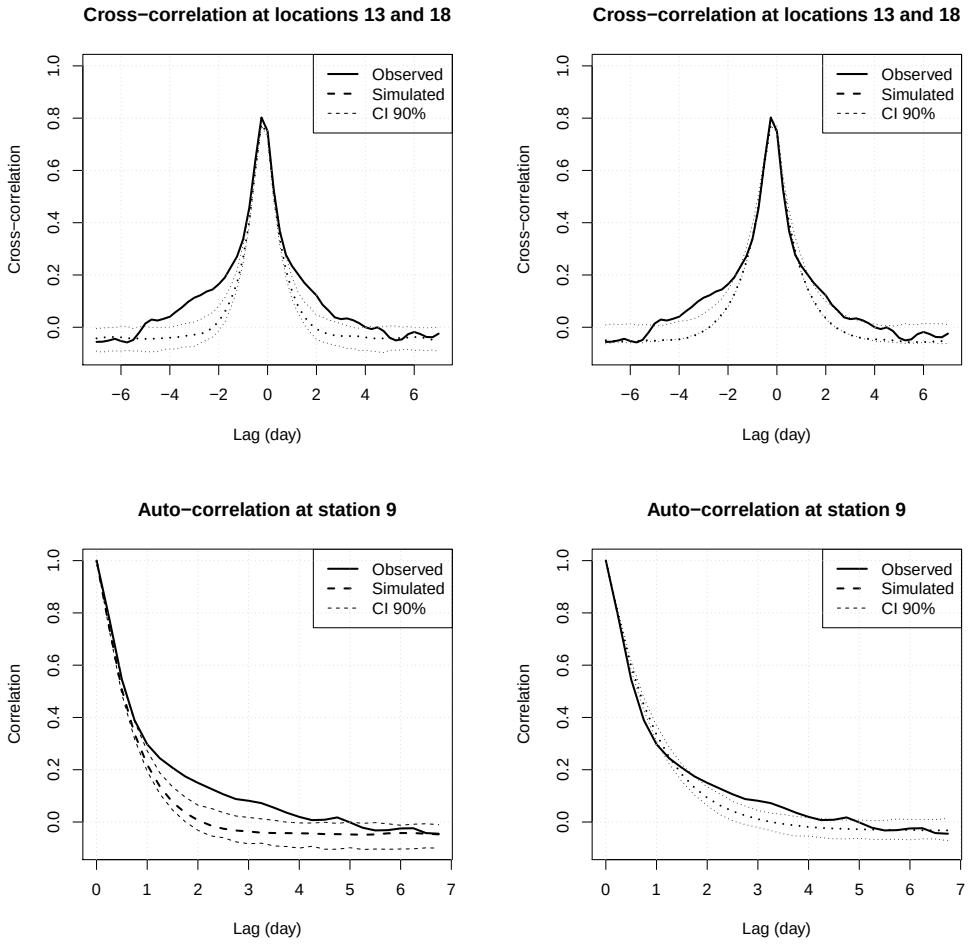

Figure 6: Observed (full lines) and simulated (dashed lines) cross-correlations between locations 13 and 18 (upper row) and auto-correlation at location 9 (lower row) for the model (M) with parameters estimated by GMM (left) and ML (right). 90% prediction intervals are computed from 100 independent simulated samples of the size of the original data.

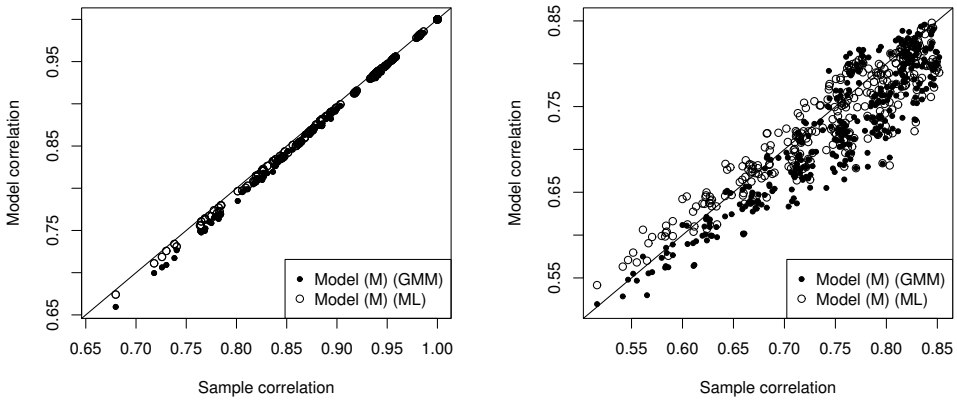

Figure 7: Theoretical correlations against observed correlations at lag 0 (left) and lag 1 (right) for the model (M) and the two methods of estimation.

5.1 Parameterization of Γ

The spatial structure of the estimated $\mathbf{\Gamma}$ shown on Figure 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 well defined covariance functions (Cressie, 1991; Abrahamsen, 1997) and

are denoted respectively ($M_{\mathbf{\Gamma}\sim Gauss}$) and ($M_{\mathbf{\Gamma}\sim Sinus}$) hereafter.

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

$$d_{i,j} = \sqrt{\Delta\text{Lat}(i,j)^2 + \theta_1 \Delta\text{Long}(i,j)^2 + \theta_2 \Delta\text{Lat}(i,j)\Delta\text{Long}(i,j)}$$

where $\Delta\text{Lat}(i,j)$ and $\Delta\text{Long}(i,j)$ 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 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. As mentioned in Section 4.1 these blocks have a particular elliptical shape that is difficult to reproduce with 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.

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) that leads to an efficient numerical procedure. The

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

Table 2: Table of log-likelihoods and BIC indexes for the different models.

models have been validated in the same way as the model (M) (see Section 4). Similar results were obtained for the marginal distributions and the temporal correlation functions but the description of the spatial structure was deteriorated when using a $(\mathrm{M}_{\mathbf{\Gamma}})$ model instead of (M). This miss-specification is also confirmed by the Bayes Information Criterion (BIC) values given in Table 2 where $\mathrm{BIC} = -2\log \mathrm{L} + \mathrm{N}_p \log(\mathrm{N}_{obs})$ with L the likelihood of the model, N_p the number of parameters and N_{obs} the number of observations. The reduced models $(\mathrm{M}_{\mathbf{\Gamma}})$ are clearly outperformed 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 that can reproduce all the complexity of the covariance matrix $\mathbf{\Gamma}$ of the observation error.

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 8). The following parameterization is then proposed

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

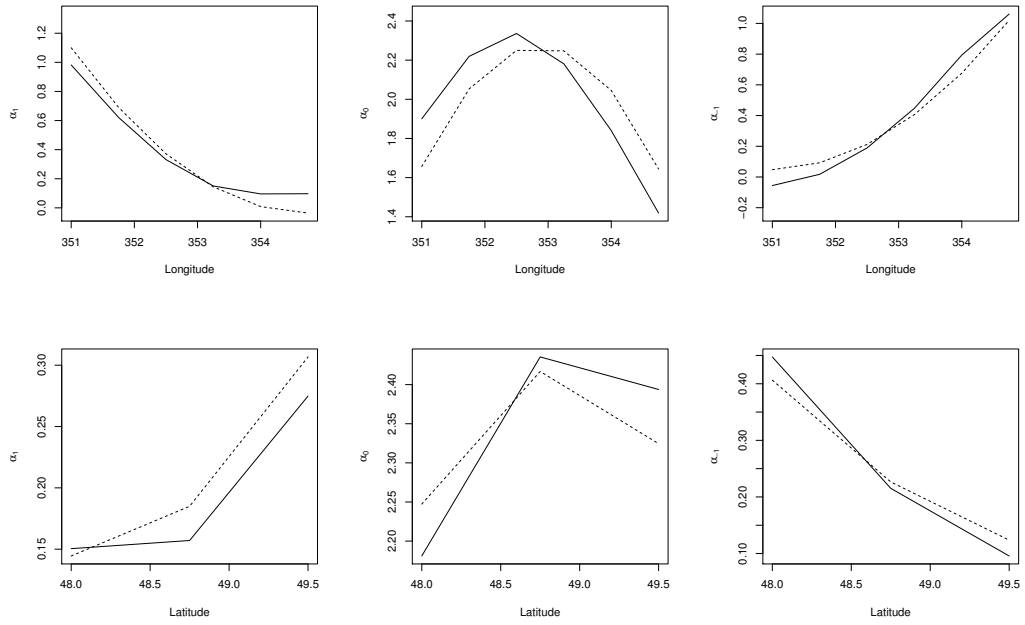

Figure 8: Estimated $\boldsymbol{\alpha}_1$ (top), $\boldsymbol{\alpha}_0$ (middle) and $\boldsymbol{\alpha}_{-1}$ (bottom) against longitude at latitude 48° N (left) and against latitude at longitude 6.75° W (right). Solid line: ML estimation of $\boldsymbol{\Lambda}$ for model (M) , dashed line: parametric structure fitted by least square.

where β_i^{Lat} for $i \in \{1, 4, 7\}$ takes a different value for each latitude and $\text{Long} \in \mathbb{R}^K$ is a vector containing the longitude of each site. Let ($\text{M}_{\boldsymbol{\Lambda}}$) denote the corresponding model. The rank of $\boldsymbol{\Lambda}$ is 3, and thus the parameters are identifiable (see Section 3.2), indeed the matrix $\begin{pmatrix} \beta_1^{\text{Lat}} & \beta_4^{\text{Lat}} & \beta_7^{\text{Lat}} \\ \beta_2 & \beta_5 & \beta_8 \\ \beta_3 & \beta_6 & \beta_9 \end{pmatrix}$ is full ranked because the matrix $\begin{pmatrix} 1 & | & \text{Long} & | & \text{Long}^2 \end{pmatrix}$ is full ranked.

The parameterization is easily handled in the GMM procedure whereas a numerical optimization is again needed to update $\boldsymbol{\Lambda}$ in the M-step of the ML procedure. Moreover a joint optimization on $\boldsymbol{\Lambda}$ and $\boldsymbol{\Gamma}$ should be done since both of them are involved in the same part of the log-likelihood function. In

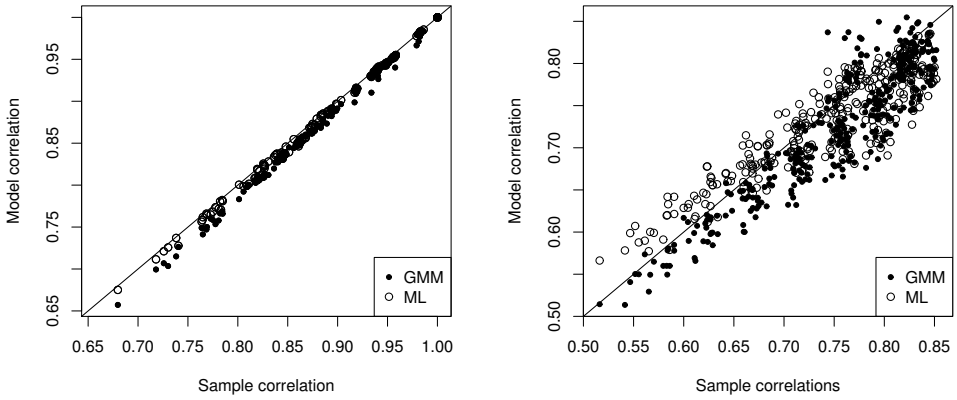

Figure 9: Theoretical correlations against observed ones at lag 0 (left) and lag 1 (right) for the model ($M_{\mathbf{\Lambda}}$).

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 ($M_{\mathbf{\Lambda}}$) and the full (M) models give similar results for the marginal distribution and the autocorrelation function. ($M_{\mathbf{\Lambda}}$) leads also to an accurate description of the spatial structure of the data (see Figure 9). Again, lagged-one correlations are better reproduced by GMM parameters than by ML parameters. The model ($M_{\mathbf{\Lambda}}$) is slightly inferior to the full model (M) in terms of BIC according to Table 2. Nevertheless, it clearly outperforms the models ($M_{\mathbf{\Gamma}}$). It seems easier to find an appropriate reduced model for the loading matrix $\mathbf{\Lambda}$ than for the covariance matrix of the observation error $\mathbf{\Gamma}$.

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 that 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 the 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 moments (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.

According to the BIC values given in Table 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. 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). In order to check the relevance of the BIC criterion, we have performed a cross-validation study (see supplementary materials) that 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.

Supplementary materials

This file contains the proof of Proposition 1, a description of the Expectation-Maximization (EM) algorithms used to fit the models introduced in the paper together with a description of the Kalman recursions involved in the EM-algorithm (supp_estimation.pdf).

Acknowledgments

We thank Peter Thomson, Statistics Research Associates, for stimulating discussions and comments on the modeling and on identifiability problems. We thank the reviewers for their helpful comments.

References

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

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.

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

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

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.

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.

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.

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.

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

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

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.

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.

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).

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.

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.

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.

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.

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.

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.

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.

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).