Automatic Classification of Offshore Wind Regimes with Weather Radar Observations

Pierre-Julien Trombe, Pierre Pinson, *Member, IEEE* and Henrik Madsen

***Abstract*—Weather radar observations are called to play an important role in offshore wind energy. In particular, they can enable the monitoring of weather conditions in the vicinity of large-scale offshore wind farms and thereby notify the arrival of precipitation systems associated with severe wind fluctuations. The information they provide could then be integrated into an advanced prediction system for improving offshore wind power predictability and controllability.**

In this paper, we address the automatic classification of offshore wind regimes (i.e., wind fluctuations with specific frequency and amplitude) using reflectivity observations from a single weather radar system. A categorical sequence of most likely wind regimes is estimated from a wind speed time series by combining a Markov-Switching model and a global decoding technique, the Viterbi algorithm. In parallel, attributes of precipitation systems are extracted from weather radar images. These attributes describe the global intensity, spatial continuity and motion of precipitation echoes on the images. Finally, a CART classification tree is used to find the broad relationships between precipitation attributes and wind regimes.

***Index Terms*—Weather radar, wind variability, Markov-Switching model, classification tree, wind energy, offshore, Horns Rev.**

I. Introduction

UNLIKE fossil fuels or nuclear energy, the availability of renewable sources of energy (e.g., solar, hydro, wind power) is directly governed by the dynamics of the atmosphere. It is therefore important to monitor weather conditions for assessing, forecasting and integrating these resources into power systems. In that respect, remote sensing observations of the atmosphere have become essential for the management of energy systems and, in offshore wind energy, they have already led to significant advances in a wide range of applications. These applications include the use of satellite SAR images for improving the accuracy of wind maps over coastal areas, airborne SAR measurements for studying wake effects at large offshore wind farms, and LiDAR and SoDAR measurements for sampling vertical wind profiles (see [1] and references therein).

A new application of remote sensing tools in wind energy is now under experimentation at Horns Rev, in the North Sea. It consists of using weather radar observations for monitoring weather conditions in the vicinity of large-scale offshore wind farms [2]. This application is motivated by the need to improve offshore wind power predictability at high temporal resolutions [3]. In particular, the high variability of offshore wind fluctuations is a serious problem for wind farm and transmission system operators because it increases the uncertainty associated with the short-term prediction of wind power [4]. Statistical analysis of wind data from Horns Rev showed that this variability was actually the result of frequent and sudden changes of wind regimes (i.e., wind fluctuations with specific frequency and amplitude) over waters [5], [6]. Subsequent analysis showed that large wind fluctuations tended to be coupled with specific climatological patterns and, particularly, the occurrence of precipitation [7]. This suggests that precipitation could be used as an early indicator for high wind variability. Our idea is thus to take advantage of the extended visibility provided by weather radars for notifying the arrival of precipitation systems in the vicinity of offshore wind farms, and adapting the forecasting strategy accordingly.

In view of integrating weather radar observations into wind power prediction systems, it is necessary to understand the precipitation settings associated with high wind variability at offshore sites. In some other meteorological contexts, the settings favoring the development of severe weather with the formation of precipitation are well documented [8], [9]. However, no detailed precipitation climatology over the North Sea exists to our knowledge. As a first step towards this understanding, we start by analyzing precipitation over the largest spatial scale enabled by the weather radar system used for monitoring the weather at Horns Rev, that is within a window of radius 240 km. Weather radar observations show that the passage of some meteorological phenomena producing precipitation was coupled with severe wind fluctuations while that of some other phenomena, also producing precipitation, was not [2]. Capturing the differences between precipitation systems by "eye" becomes increasingly difficult with the volume of data. This difficulty may further be increased by other factors such as (i) the relatively small range of single weather radar systems which only enables a partial observation of precipitation systems; (ii) seasonal variations of precipitation which implies that two similar events on weather radar images at two different times of the year may have different levels of severity. This calls for the use of statistical classifiers for generating a consistent catalogue of situations where the variability of wind fluctuations is explained by attributes (i.e., characteristics) of precipitation systems.

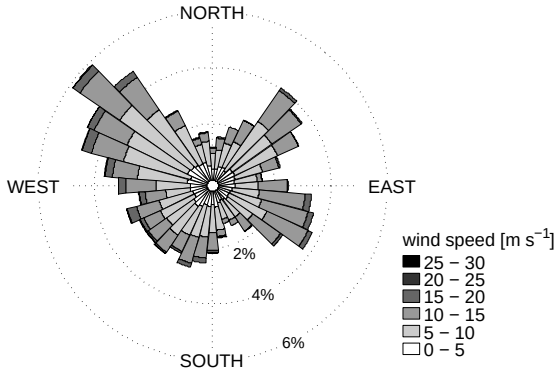
Traditionally, classification applications using precipitation attributes aim at improving the understanding of precipitation itself. For instance, an automated classification procedure for rainfall systems is proposed in [10]. Alternatively, [11], [12] address the classification of precipitation objects (i.e., storms) that require to be defined and identified a priori. Yet, a major drawback of these approaches is that they rely on an expert training performed manually with its inherent shortcomings: (i) the potential lack of consistency since two experts may

The authors are with the Department of Informatics, Technical University of Denmark, Kgs. Lyngby, Denmark. (e-mail: {pjt, pp, hm}@imm.dtu.dk)

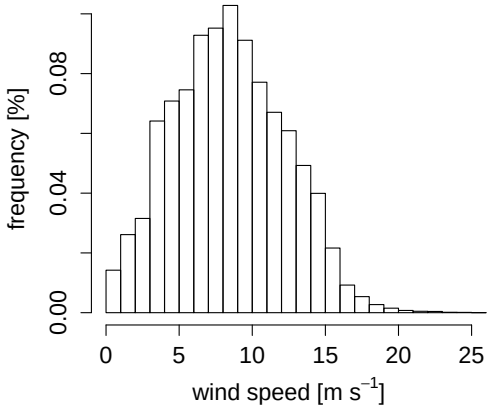
Manuscript received month day, year; revised month day, year.

disagree on how to classify an event, or a same expert may classify two similar events differently; (ii) it is limited in the volume of data that can be treated. Our study differs in two aspects. First, the target variable is not precipitation but wind. And second, it does not require any expert training for the classification and therefore avoid the aforementioned shortcomings. Instead, a categorical sequence of wind regimes is automatically estimated from a wind speed time series by combining a global decoding algorithm, the Viterbi algorithm [13], with the Markov-Switching model proposed in [5]. In parallel, a number of precipitation attributes are computed from weather radar images. These attributes describe the global intensity, spatial continuity and motion of precipitation echoes on the images. Finally, a CART classification tree, is used for finding relationships between precipitation attributes and wind regimes observed at Horns Rev. The motivation for using such a classification technique is that it can explore large amounts of data and, yet, produce a simple partition with interpretable rules [14].

The rest of the paper is organized as follows. In Section II, we describe the data. In Section III, we give an overview of the procedure for extracting the most likely sequence of regimes from wind speed time series. In Section IV, we compute a number of precipitation attributes from weather radar images. In Section V, we present the classification tree technique and apply it to the problem of the automatic classification of offshore wind regimes. Finally, Section VI delivers concluding remarks.



(a) Wind rose. Angles indicate the direction from which the wind blows (meteorological conventions).



(b) Frequency histogram of wind speed.

Fig. 1. Wind distribution at the Horns Rev 1 wind farm, in 2010. Data were collected from the nacelle anemometry and SCADA systems [15].

II. Data

A. Wind data

Wind data were collected from the nacelle anemometry and SCADA systems of the Horns Rev (HR1) wind farm [15]. The original measurements consisted of individual time series of wind speed and wind direction, for each of the 80 wind turbines of HR1. Two aggregated time series of wind speed and wind direction were obtained by jointly averaging these individual time series over 10 minute intervals. The time series span the year 2010. Due to some technical problems, measurements are missing over 2 periods of about 5 and 12 days, respectively. There are 2664 missing values out of 52560 (i.e., 94.9% of data availability). No attempt was made to fill in those gaps. The wind distribution is shown in Figure 1. The wind rose shows 3 preferred wind directions. While the prevalence of northwesterly directions is consistent with other wind data analysis at Horns Rev (see [6]), the frequent occurrences of northeasterly winds are more exceptional since it is usually the direction where the wind is suppressed in Denmark. This phenomenon can be explained by a strong annual wind variability in 2010. Note also that strong winds, above 15 m s^{-1}, are more frequent for westerly than easterly directions.

B. Weather radar data

Weather radar data consist of 2D images of precipitation reflectivity. More specifically, they correspond to 1 km height pseudo-CAPPI (Constant Altitude Plan Position Indicator) image products, with a 2×2 km grid resolution. They were produced by a C-Band Doppler radar located in Rømø, approximately 57 km to the East of the HR1 wind farm. The radar is operated by the Danish Meteorological Institute (DMI), using a 9 elevation scan strategy and an operational range of 240 km [16]. One image is generated every 10 minutes. Clutter removal filters are applied during the data acquisition process. Data quality control is also performed a posteriori and persistent clutter is removed following the automatic method introduced in [17]. For a complete description of the radar settings and images, we refer to [2]. About 2000 images are missing over the year 2010 (i.e., 96.1% of data availability).

III. Estimation of wind regimes

In this section, we estimate a categorical sequence of wind regimes from the time series of wind speed presented in Section II. Such a procedure can also be viewed as a segmentation of the time series where the latter is partitioned into homogeneous sections. Our plan is to use this sequence of wind regimes as the dependent variable (i.e., the variable to predict) for growing a classification tree in Section V.

Numerous studies have pointed out the nonstationary behavior of offshore or near-offshore wind fluctuations at the minute scale [6], [18], [19]. Numerically, this nonstationarity translates into sudden shifts in the amplitude and/or frequency of wind fluctuations. Such patterns of fluctuations can be

analyzed either in the frequency domain, with an empirical spectral decomposition technique as in [6], or in the time domain with Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) models [18], or Markov-Switching AutoRegressive (MSAR) models [5]. The advantage of MSAR models over other techniques is that they are clearly tailored to address the extraction of a hidden sequence of regimes, as discussed in [20].

We follow a 2-step procedure. First, a MSAR model is fitted to the time series of wind speed. Secondly, a global decoding method, the Viterbi algorithm [13], is used for computing the most likely sequence of wind speed regimes, under the fitted MSAR model.

A. Regime-switching modeling with MSAR models

MSAR models are an extension of Hidden Markov Models (HMM). They are widely used for the modeling of time series characterized by structural breaks in their dynamics. The underlying assumption of these models, both HMM and MSAR, is that there is an unobservable Markov process which governs the distribution of the observations [20]. Compared to HMM, MSAR models have an additional capability, they can accommodate autocorrelated data and include autoregressors in the model formulation. Applications of MSAR models to wind data include [5], [21].

The wind speed time series we use for this study does not show any well pronounced diurnal cycle. In addition, we disregard the potential long-term drift and seasonal variations of wind speed since the available time series only spans a one year period. For the sake of simplicity, we do not specifically deal with the wind speed truncation in 0. We only assume that wind speed has an autoregressive behavior in each regime. Let $\{y_t\}, t = 1, \ldots, n$, be the time series of measured wind speed at the HR1 wind farm. The MSAR model with m regimes and autoregressive orders $(p_1, \ldots, p_m)$ is defined as follows:

$$Y_t = \boldsymbol{\theta}^{(Z_t)T} \boldsymbol{X}_t + \sigma^{(Z_t)} \varepsilon_t^{(Z_t)} \tag{1}$$

with

$$\boldsymbol{\theta}^{(Z_t)} = [\theta_1^{(Z_t)} \quad \ldots \quad \theta_{p_{Z_t}}^{(Z_t)}]^T \tag{2}$$

$$\boldsymbol{X}_t = [Y_{t-1} \quad \ldots \quad Y_{t-p_{Z_t}}]^T \tag{3}$$

where $\{\varepsilon_t\}$ is a sequence of independently distributed random variables following a Normal distribution $\mathcal{N}(0,1)$; and $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ is a first order Markov chain with a discrete and finite number of states (i.e., regimes) m and transition probability matrix $\boldsymbol{P}$ of elements $(p_{ij})_{i,j=1,\ldots,m}$ with:

$$p_{ij} = Pr(Z_t = j | Z_{t-1} = i), \quad i, j = 1, \ldots, m \tag{4}$$

$$\sum_{j=1}^{m} p_{ij} = 1, \quad i = 1, \ldots, m \tag{5}$$

There exist two distinct methods for estimating the parameters of a MSAR model with given number of regimes m and autoregressive orders $(p_1, \ldots, p_m)$, the Expectation-Maximization (EM) algorithm and direct numerical maximization of the Likelihood. The respective merits of these 2 methods are discussed in [20], along with practical solutions for their implementation. As for this study, we estimate MSAR models by direct numerical maximization of the Likelihood owing to its lower sensitivity to starting values. Let $\boldsymbol{\Theta} = (\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(m)}, \boldsymbol{P}, \boldsymbol{\sigma})$ be the set of parameters to estimate. The Maximum Likelihood Estimator (MLE), $\widehat{\boldsymbol{\Theta}}_{MLE}$, is obtained by maximizing the Likelihood function $L(\boldsymbol{\Theta})$:

$$\widehat{\boldsymbol{\Theta}}_{MLE} = \arg\max_{\boldsymbol{\Theta}} L(\boldsymbol{\Theta}) \tag{6}$$

$$= \arg\max_{\boldsymbol{\Theta}} \boldsymbol{\delta} (\prod_{t=1}^{n} \boldsymbol{P} \boldsymbol{D}_t) \mathbf{1}^T \tag{7}$$

with

$$\boldsymbol{\delta} = \mathbf{1} (\boldsymbol{I}_m - \boldsymbol{P} + \boldsymbol{U}_m)^{-1} \tag{8}$$

$$\boldsymbol{D}_t = \text{diag}(\eta(t,1), \ldots, \eta(t,m)) \tag{9}$$

$$\eta(t,i) = \frac{1}{\sigma^{(i)}} \phi \left(\frac{Y_t - \boldsymbol{\theta}^{(i)T} \boldsymbol{X}_t}{\sigma^{(i)}} \right), \quad i = 1, \ldots, m \tag{10}$$

$\boldsymbol{\delta}$ is the stationary distribution of the Markov chain; $\mathbf{1}$ is a unit vector of size m; $\boldsymbol{I}_m$ and $\boldsymbol{U}_m$ the Identity and Unity matrices of size $m \times m$; $\boldsymbol{D}_t$ a diagonal matrix; and ϕ the probability density function of the Normal distribution.

We estimate four MSAR models, from one up to four regimes. For each of these MSAR models, the optimal autoregressive orders in each regime are determined by following a forward selection procedure based on Likelihood Ratio (LR) tests, as described in [22]. Then, all four models are compared with one another by performing LR tests, leading to the rejection of the MSAR model with four regimes. For MSAR models from one to three regimes, Table I summarizes some of the important parameter estimates that help interpreting the regimes. In particular, the elements of the diagonal of the transition probability matrix, diag($\boldsymbol{P}$), give an estimation of the mean persistence of the regimes over time. As for the vector of standard deviations $\boldsymbol{\sigma}$, it expresses the relative variability of wind speed fluctuations in each regime. The estimates of the autoregressive coefficients are of lesser importance and, instead, we just report the optimal autoregressive order in each regime. Regimes are ranked by ascending values of standard deviation. Both with 2 and 3 regimes, there is an inverse relationship between wind fluctuation variability and persistence (i.e., the more variable, the less persistent).

TABLE I
SUMMARY STATISTICS ON MSAR MODELS FITTED TO THE TIME SERIES OF WIND SPEED.

m	$(p_1, \ldots, p_m)$	diag($\boldsymbol{P}$)	$\boldsymbol{\sigma}$
1	5	-	0.51
2	(5,5)	(0.98, 0.92)	(0.31, 0.96)
3	(4,3,6)	(0.98, 0.95, 0.89)	(0.25, 0.47, 1.28)

B. Global decoding

Global decoding consists of estimating the most likely sequence of regimes $\widehat{\boldsymbol{z}} = (\hat{z}_1, \ldots, \hat{z}_n)$ under a fitted model, as opposed to *local decoding* which consists of estimating the

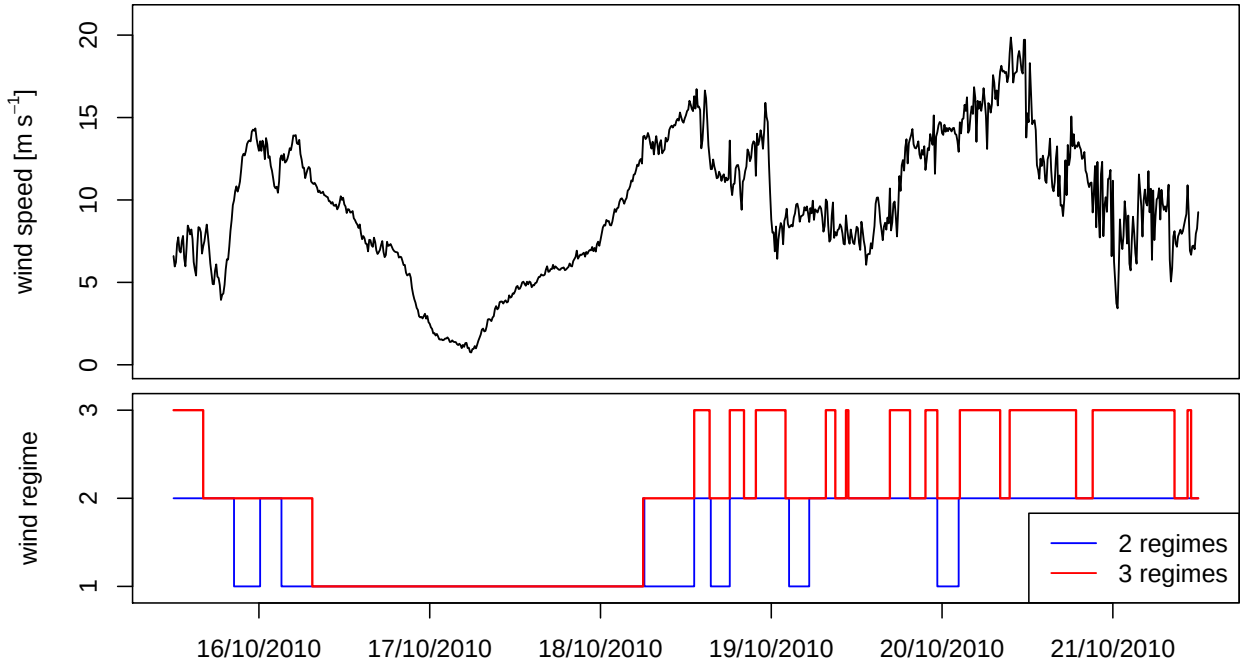


Fig. 2. (Upper panel) Time series of wind speed recorded at the Horns Rev 1 wind farm. The temporal resolution is 10 minutes. (Lower panel) Estimated sequence of regimes, for 2 and 3 regimes. Regimes can be interpreted in terms of wind variability, from low in regime 1 to high variability in regime 3.

most likely regime at time t, $\hat{z}_t$, independently of the regime values at other times. The most likely sequence of regimes $\widehat{\boldsymbol{z}}$ is found by maximizing the joint probability of the observations and states of the Markov chain:

$$\widehat{\boldsymbol{z}} = \arg\max_{\boldsymbol{z}} \Pr(\boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{Y} = \boldsymbol{y}) \tag{11}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. For estimating $\widehat{\boldsymbol{z}}$, we use the Viterbi algorithm [13]. For that purpose, let us introduce the following notations:

$$\boldsymbol{Y}^{(i)} = (Y_1, \ldots, Y_i) \quad \text{and} \quad \boldsymbol{Z}^{(i)} = (Z_1, \ldots, Z_i) \tag{12}$$

$$\xi_{1i} = \Pr(Z_1 = z_1, Y_1 = y_1) = \delta_i \eta(1, i) \tag{13}$$

$$\xi_{ti} = \max_{\boldsymbol{z}^{(t-1)}} \Pr(\boldsymbol{Z}^{(t-1)} = \boldsymbol{z}^{(t-1)}, Z_t = i, \boldsymbol{Y}^{(t-1)} = \boldsymbol{y}^{(t-1)}) \tag{14}$$

for $t = 2, \ldots, n$. The quantities ξ_{ti} can be seen as the most probable sequence leading to regime i at time t, among all possible sequences $\boldsymbol{Z}^{(t-1)}$. Finally, $\widehat{\boldsymbol{z}}$ is found by the solving the following backward recursion, starting from n:

$$\hat{z}_n = \arg\max_{i=1,\ldots,m} \xi_{ni} \tag{15}$$

$$\hat{z}_t = \arg\max_{i=1,\ldots,m} \xi_{ti} p_{i,\hat{z}_{t+1}} \quad \text{for} \quad t = n-1, \ldots, 1 \tag{16}$$

The most likely sequence of wind regimes was computed under the fitted MSAR models, with both 2 and 3 regimes. The result is illustrated in Figure 2 over a 6 day episode where a clear change of wind speed variability, from low to high, can be observed on October 18, 2010. Note that the regimes are more stable (i.e., there are fewer switchings) for the sequence with 2 regimes than with 3.

IV. Precipitation identification and attributes

A. Precipitation identification

Weather radar images can contain 2 sources of information which fall either into the meteorological sources (e.g., rain, hail, snow) or into non-meteorological sources (e.g., clutter due to buildings, wind farms, ground, sea). Echoes caused by non-meteorological targets can usually be identified and filtered out during the data acquisition process or a posteriori data quality control when they have non-random patterns (see [23] for illustrative examples on the Danish weather radar networks). However, not all non-meteorological echoes can be removed and, in some cases, significant portions of weather radar images remain contaminated by non-meteorological artifacts [2]. Regarding the images used in this study, the most serious problems are due to anomalous propagation (anaprop) of the radar beam. We observe these problems more frequently during the summer season, from April to September in Denmark. In some extreme cases, the contamination can extend up to 20% of the image pixels over several hours. Image pre-processing operations such as median filtering are inefficient for removing anaprop echoes.

In this subsection, our goal is to develop a method for assigning a binary label to each image indicating the detection of precipitation (potentially mixed with noisy echoes) or not. In [24], rainfall is identified by computing the proportion of wet pixels (i.e., pixels recording positive rainfall) over the entire image. A rainfall event is then defined as a continuous period of time where the coverage proportion of wet pixels over the whole image is above a threshold of 25%. This approach is clearly an over-simplified view of the problem and could not apply to our images, even by optimizing the threshold level. In other applications and, particularly, severe

weather nowcasting, storm identification is addressed by defining thresholding and contiguity heuristics [25]. These later methods are tailored for very specific types of precipitation being depicted by high reflectivity echoes on weather radar images.

We propose an alternative method for identifying precipitation, irrespectively of the mean reflectivity. It is based on the assumption that contiguous pixels recording precipitation have a higher correlation than contiguous pixels contaminated by noise. This assumption is supported by [26] which shows that the shape of precipitation echoes tends to be elliptical. We use a geostatistical tool, the correlogram, as a measure of spatial correlation of precipitation echoes for each image [27]. In order to capture the potential anisotropy of precipitation echoes, these correlograms are produced in 2 dimensions, based on the estimation of directional correlograms $\rho(\boldsymbol{h})$ of vector $\boldsymbol{h}$ as follows:

$$\rho(\boldsymbol{h}) = \frac{\gamma(\boldsymbol{h})}{\gamma(0)} \tag{17}$$

$$\gamma(\boldsymbol{h}) = \frac{1}{N(\boldsymbol{h})} \sum_{(p_i,p_j)|h_{p_i p_j}=\boldsymbol{h}} (I_{p_i} - I_{p_j})^2 \tag{18}$$

where $\gamma(\boldsymbol{h})$ is a directional variogram computed by summing over all paired pixels (p_i, p_j) with intensities (I_{p_i}, I_{p_j}) and separated by a vector $\boldsymbol{h}$. $N(\boldsymbol{h})$ is the number of paired pixels (p_i, p_j) matching this latter criterion. These 2-dimensional correlograms are computed with the *gstat* package of the R programming environment [28].

Figure 3 shows 4 sample images and their associated correlograms. A zoom in the central part of the correlogram is also provided for illustrating the local continuity of reflectivity values. The images were chosen to reflect various types of precipitation systems (e.g., small and scattered precipitation cells, banded or widespread precipitation system) and a case of anaprop. In particular, the small spatial correlation of anaprop echoes can well be observed, it drops below 0.4 for all 1-lagged (i.e., adjacent) pixels, whatever the direction. Note also the quick decorrelation in space for small scattered cells but, unlike for anaprop, the spatial correlation is larger than 0.4 up to 3-4 lagged pixels. The anisotropy of banded systems can also be well be captured by these correlograms.

For a given image, we consider that precipitation is detected if the correlation is larger than 0.6 for all 1 and 2-lagged pixels (i.e., the central 5x5 neighborhood of the correlogram). Then, we define a precipitation event as a period with a minimum duration of 1 hour (i.e., 6 consecutive images) over which precipitation is detected. If the time between the end of a precipitation event and the beginning of a new one is less than one hour, we consider it to be the same event. Table II summarizes the number of events identified and their mean lifetime in 2010.

B. Precipitation types

Precipitation is commonly described as either stratiform, convective or a mix of these two. In the mid-latitudes, stratiform precipitation develops in a variety of situations where the atmosphere is stably stratified. Typical examples of these situations are warm fronts where masses of warm air gradually lift over cold masses of air. These fronts have the particularity of propagating relatively slowly and spreading over large horizontal scales up to and beyond 100 km. On weather radar images in 2D, stratiform precipitation is thus generally identified as a widespread region of moderate, homogeneous and continuous intensity with a slow dynamics. Winds associated with pure stratiform precipitation usually have a small vertical velocity and low turbulency. In comparison, convective precipitation develops in unstable atmosphere and have a much higher spatial variability, with many scattered and heavy precipitation showers occurring locally, over horizontal scales from a few kilometers up a few tens of kilometers, potentially forming complex convective systems over several hundreds kilometers. In addition, the updraft associated with this type of precipitation is stronger, resulting in highly turbulent winds. In the mid-latitudes, convective precipitation prevails during the summer and over warm oceans. On weather radar images in 2D, convective precipitation is depicted by small clusters of high reflectivity propagating relatively quickly. However, in many cases, convective precipitation can be embedded into stratiform regions and forms more complex precipitation structures.

C. Precipitation attributes

For each image where precipitation is identified, we compute a number of attributes linked the global intensity, spatial continuity and motion of precipitation. These attributes are meant to describe the main characteristics of the different types of precipitation discussed hereabove. They are summarized in Table III.

On weather radar images, the intensity of precipitation is measured in decibel of reflectivity (dBZ). Within a same precipitation system, the distribution of intensity may not be homogeneous and, with the occurrence of severe weather traditionally associated with high values of reflectivity, it tends to be positively skewed. So, in order to describe the distribution of precipitation intensity, we propose a set of non parametric statistics composed of (i) location measures with the median (i.e., the 50^{th} quantile), the 75^{th}, 90^{th}, 95^{th} and 99^{th} quantiles; (ii) dispersion measures with the interquartile range (i.e., the range between the 25^{th} and 75^{th} quantiles); (iii) shape measures with the skewness to inform on the asymmetry of the distribution, and the kurtosis to inform on its sharpness. Only pixels with strictly positive reflectivity values are considered. Note that we choose to use robust statistics with, for instance, the median in place of the mean and the 99^{th} quantile in place of the maximum in order to filter out the potential effects of residual noise.

For measuring the spatial continuity of precipitation, we again use the correlogram introduced in this Section and follow the procedure presented in [10]. It assumes that each correlogram contains an elliptical object that can be described by its eccentricity and area. The procedure is as follows: (1) the correlogram is transformed into a binary image by means of a thresholding operation, with the threshold value arbitrarily chosen between 0 and 1; (2) a connected-component labelling

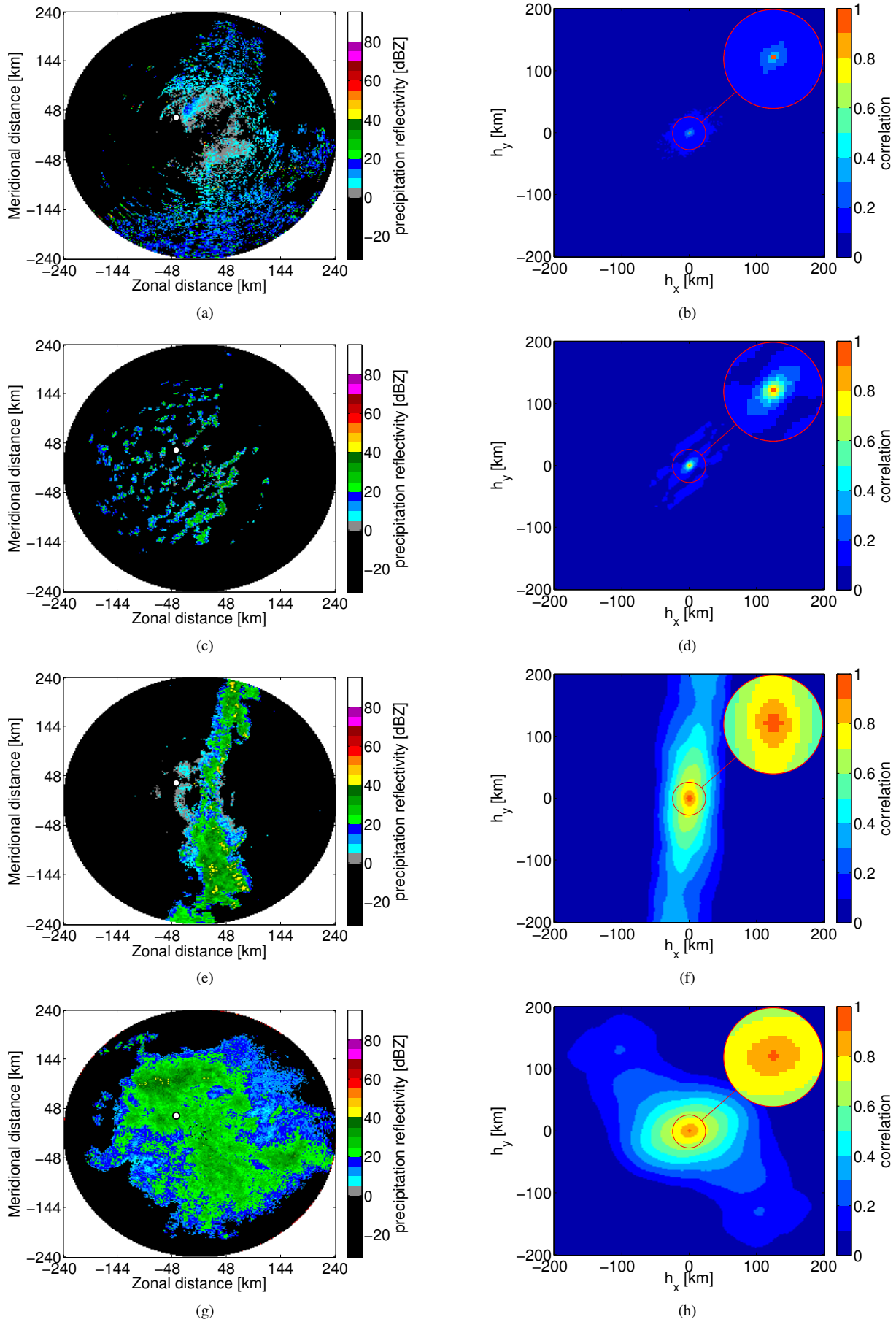


Fig. 3. Image samples (left column) and their associated correlograms in 2 dimensions (right column). (a-b) A case of anomalous propagation without precipitation. (c-d) Small scattered convective precipitation cells. (e-f) Banded precipitation system. (g-h) Widespread precipitation system.

TABLE II
MONTHLY STATISTICS: NUMBER OF PRECIPITATION EVENTS IN 2010 AND THEIR MEAN LIFETIME

	Jan.	Feb.	March	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Total
Number of events	23	19	40	20	32	18	24	25	20	24	36	21	302
Mean lifetime [hours]	22.7	24.2	08.6	14.9	11.1	12.3	20.4	21.9	20.5	21.2	16.2	20.4	17.1

algorithm is used to identify all connected regions on the binary image [29] and only the region intersecting with the center of the image is kept; (3) the edge of that region is identified with the Canny edge detector [30]; (4) an ellipse is fitted on the detected edge by minimizing the least square criterion [31]. In this study, this procedure is performed twice, for threshold values of 0.4 and 0.7, and the eccentricity (i.e., the ratio of the major axis over the minor axis) and the area of the elliptical object are computed for both values. For the threshold value of 0.4, these attributes are likely to reflect the large-scale continuity of precipitation whereas, for the value of 0.7, they will capture the more local continuity.

The horizontal motion of precipitation is computed with an optical flow method. This type of method is very useful for estimating the visible flow field (u, v) between 2 consecutive images. The two underlying assumptions that define the optical flow formulation are *brightness constancy* and *spatial smoothness*. Brightness constancy means that the intensity of an object is conserved over time, despite its likely change of position between 2 consecutive images. Spatial smoothness refers to the coherence between neighboring pixels which should ideally have a similar motion [32]. The formulation we use is the one Black and Anandan proposed in [33] owing to its robustness to outliers. It is expressed as an Energy minimization problem with the objective function $E(u, v)$ defined as follows:

$$E(u, v) = E_{BC}(u, v) + \lambda E_{SS}(u, v) \tag{19}$$

where λ a regularization parameter (i.e., the larger λ, the denser the flow field); E_{BC} and E_{SS} are the functions resulting from the brightness constancy and spatial smoothness constraints:

$$E_{BC}(u, v) = \sum_{i,j} f(I_1(i, j) - I_2(i + u_{i,j}, j + v_{i,j})) \tag{20}$$

$$E_{SS}(u, v) = \sum_{i,j} [g(u_{i,j} - u_{i+1,j}) + g(u_{i,j} - u_{i,j+1}) + g(v_{i,j} - v_{i+1,j}) + g(v_{i,j} - v_{i,j+1})] \tag{21}$$

where I_1 and I_2 are 2 consecutive images, f and g are 2 penalty functions. Following the implementation of Black and Anandan, we set $f = g = log(1 + \frac{1}{2}(\frac{x}{\sigma})^2)$, the Laurentzian function with scale parameter σ. The expression of E_{SS} is formulated with a pairwise Markov Random Field (MRF) discretization, based on a 4-neighborhood [34]. Since our goal is to estimate a unique speed and direction for each pair of consecutive images, we extract the median Cartesian flow from the flow field and convert it into its Polar components (i.e., speed and direction). Flow direction is then transformed into a categorical variable by binning its values into 8 sectors (North (N), North-East (NE), ...).

Finally, we also add a seasonal attribute in the form of a categorical variable to allow for potential seasonal patterns of precipitation. We consider that there are only two seasons in Denmark so that the variable takes value *Summer* from March to August, and *Winter* from September to February. In summer, the North Sea is on average colder than the air whereas, in winter, the opposite holds true and favors thermal instabilities in the atmosphere [7].

V. AUTOMATIC CLASSIFICATION

For the automatic classification of precipitation systems, we use a tree-based classification technique called CART, in a supervised learning framework (i.e., the classification is governed by the categorical sequence of wind regimes computed in Section III). These trees, also known as decision trees, are attractive in many aspects. First, for the relative simplicity of their principles based on a recursive partitioning of the data set. Second, they provide a powerful alternative to more traditional classification techniques (e.g., discriminant analysis and logistic regression) which generate a global model for the entire data set while variables may interact in a highly complex and nonlinear way and require to be fitted locally. Finally, because their interpretation is mainly visual and can lead to a straightforward understanding of the relationships between variables [14]. Applications of classification trees to precipitation data extracted from weather radar images can be found in [11], [12].

A. CART classification trees

Let Y be the dependent categorical variable taking values $1, 2, \ldots, K$, and $(X_1, \ldots, X_p)$ the set of p predictors (i.e., the independent variables) that can either be continuous or categorical. Growing a classification tree consists of a recursive partitioning of the feature space (i.e., the space composed of the p predictors each with n observations) into rectangular areas. Each split consists of a dichotomy applied on a single predictor (e.g., $X_2 < 3$ if X_2 is continuous or $X_2 = "a"$ if it is categorical). The feature space is first split into 2 groups so that the response of Y is maximized in each of the 2 groups. This procedure is recursively repeated and each of the 2 groups is partitioned into 2 new sub-groups, and so on. Splits are more commonly called *nodes*. A terminal node (i.e., node that cannot be further split) is called a *leaf*.

For each node, the splitting predictor and rule are determined so as to minimize the impurity level in the resulting two nodes. For a given node, let $p = (p_1, \ldots, p_K)$ be the vector of proportions of elements in class $1, \ldots, K$. There exist several impurity measures and the one we use in this study is known as the Gini index. It measures how often a randomly chosen

TABLE III
DESCRIPTION OF PRECIPITATION ATTRIBUTES USED FOR GROWING THE CLASSIFICATION TREE.

Attribute acronyms	Type (source)	Unit	Description
skew & kurt	Intensity (reflectivity images)	-	Skewness and Kurtosis of reflectivity distribution
q50, q75, q90, q95 & q99	Intensity (reflectivity images)	dBZ	50^{th}, 75^{th}, 90^{th}, 95^{th} & 99^{th} reflectivity quantiles
iqr	Intensity (reflectivity images)	dBZ	Interquartile range (range defined by the 25^{th} and 75^{th} reflectivity quantiles)
speedOF	Motion (optical flow)	m s^{-1}	Median speed of the flow field
dirOF	Motion (optical flow)	N, NE, E, SE, S, SW, W, NW	Median direction (8 sectors) of the flow field. Direction are in meteorological conventions, they indicate the direction of origin.
spaArea04, spaArea07	Spatial continuity (correlogram)	km^2	Area of the ellipse fitted on correlograms for threshold values 0.4 and 0.7
spaEcc04, spaEcc07	Spatial continuity (correlogram)	-	Eccentricity of the ellipse fitted on correlograms for threshold values 0.4 and 0.7
season	Temporal	Sm./Wt.	Summer (from April to September), Winter (from October to March)

element from the node would be incorrectly labeled if it were labeled according to the frequency distribution of labels in the node. The Gini index $i_G(p)$ is computed as follows:

$$i_G(p) = 1 - \sum_{j=1}^{K} p_j^2 \tag{22}$$

When growing a tree, the tradition is to build a complex tree and simplify it by pruning (i.e., removing the nodes that over-fit the feature space). This is done by minimizing the misclassification rate within leaves over a 10-fold cross-validation procedure.

B. Experimental results

The classification is performed using the sequence of wind regimes computed in Section III as the dependent variable, and the precipitation attributes extracted from the weather radar images and listed in Table III as predictors. Observations where no precipitation is detected are filtered out. After that, more than 29000 observations remain for the classification. We choose to grow the tree for the sequence of wind regimes with 2 regimes. There are 76% of observations in regime 1 and 24% in regime 2. The final tree is shown in Figure 4. Branches going downwards to the left indicate that the splitting rule is satisfied.

The classification tree we grew is interesting in two aspects. First, it reveals the broad patterns of precipitation systems associated with the different wind regimes. For instance, the leftmost leaf which contains 35% of the total number of observations, shows that 93% of the observations for which the speed of precipitation echoes is smaller than 12 m s^{-1} (i.e., speedOF$<$12) and the maximum reflectivity is smaller than 29 dBZ (i.e., reflQ99$<$29) are in Regime 1. On the opposite side of the tree, the rightmost leaf which contains 14% of the total number of observations, indicates that 59% observations for which the speed is larger than 12 m s^{-1}, the maximum reflectivity larger than 30 dBZ and the precipitation comes from North-West, West or South are in Regime 2. One recurrent pattern in this tree is that when precipitation systems comes from North-East, East or South-East, wind fluctuations tend be classified in Regime 1, the regime with the lowest variability. This is consistent with the results in [7] that show that wind fluctuations are more variable for westerly flows than for easterly flows.

Secondly, the tree highlights the predictive power of each of the variables used in the classification. Some variables may repeatedly be used for generating new nodes whereas some other variables may not be used at all. This contrasts with the hierarchical clustering technique proposed in [10] where all variables equally contribute to classify observations, with the risk of including non informative variables and degrading the accuracy of the classification. In the present experiment, one can notice that only 4 predictors are used in the final tree, the motion speed and direction of precipitation echoes (i.e., speedOF and dirOF), the season and the maximum reflectivity (i.e., reflQ99). Note that the maximum reflectivity value (i.e., reflQ99) is the only intensity related attribute used in the final tree. This attribute characterizes the most extreme, yet marginal, intensity recorded on the images, highlighting the necessity to consider precipitation information at smaller scales in the future. Moreover, none of the 4 variables derived from the correlogram (i.e., spaArea04, spaArea07, spaEcc04 and spaEcc07) is used. The most likely reason for the small predictive power of correlograms is the too complex organizational structure of precipitation systems. In particular, when there are spatial discontinuities between precipitation echoes (i.e., precipitation echoes are separated by regions recording no precipitation), correlograms are only informative locally and cannot capture the full extent of the precipitation system. Inversely, when small clusters of high intensity are embedded into a large and continuous region of moderate intensity, correlograms tend to only capture the large-scale feature. This suggests the development of hierarchical techniques where precipitation would be analyzed at multi-scale, as a potential line of work in the future.

VI. CONCLUSION

In this work, we proposed an automatic procedure for classifying offshore wind regimes based on precipitation attributes extracted from weather radar images. We found that winds with a high variability are more likely to be observed with the passage of precipitations systems being advected at relatively high speeds, preferably from West and North-West, and having large maximum reflectivity values. This result is consistent with earlier data analysis [7] and confirms the potential of

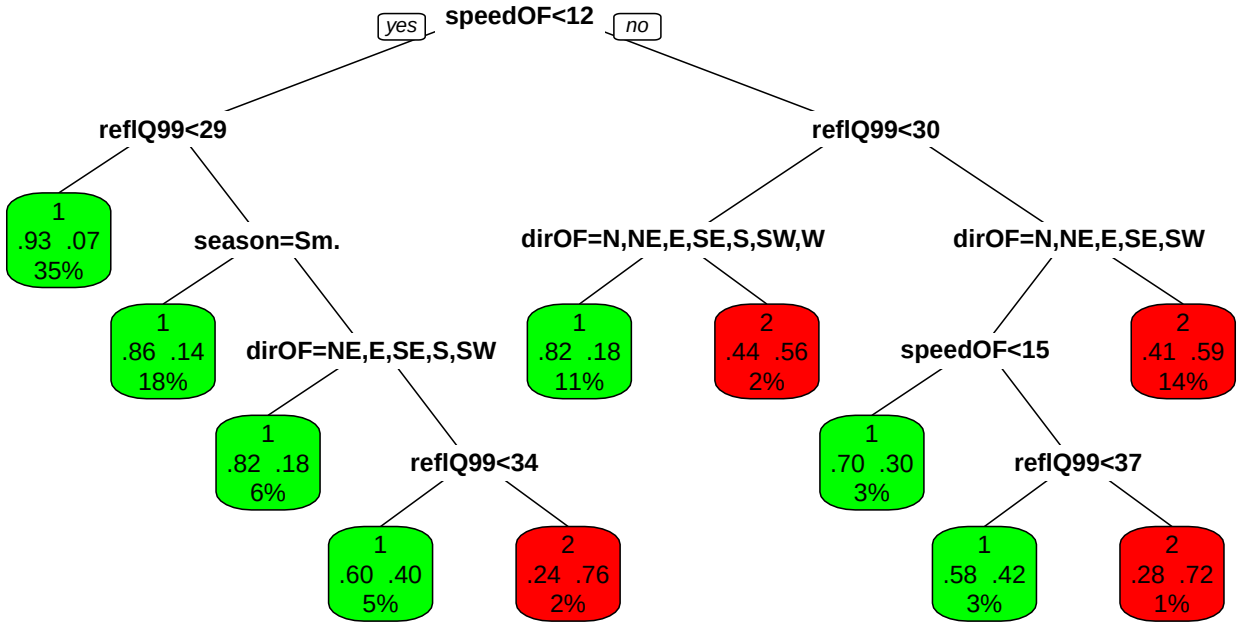


Fig. 4. Classification tree explaining wind regimes at the Horns Rev 1 wind farm with precipitation system attribute extracted from weather radar images. Wind regimes and precipitation system attributes are computed in Section IV and V, respectively.

weather radar observations for providing appropriate information to future wind power prediction systems. However, the insights we gained on the relationship between precipitation and wind are not readily integrable into prediction systems.

We studied wind fluctuations in a univariate framework, only considering wind speed. It has the merit of keeping the complexity of the procedure reasonable. Yet, wind should ideally be considered and treated as a bivariate process of speed and direction because patterns of wind speed fluctuations may either be direction-dependent or coupled with specific patterns of wind direction fluctuations. For instance, larger wind speed fluctuations are observed for westerly flows at Horns Rev [7]. However, the statistical modeling of circular time series (e.g., wind direction) that feature autocorrelation is quite cumbersome and it is preferable to transform wind speed and direction into their associated (u, v) components, as in [18], for instance. That way, both variables of the bivariate process are non-circular and unbounded, and traditional methodologies can be applied. In that view, an interesting generalization of our work could consist of applying MSAR models in a vectorial form as introduced in [35], on the transformed (u, v) components of the wind.

As for precipitation, we considered it over a unique and large spatial scale which is suitable for a preliminary investigation aiming at defining a rough climatology of precipitation and wind. However, our approach clearly overlooks the important organizational structure of precipitation systems. This acts as a limiting factor for improving the accuracy of the classification of offshore wind regimes. A potential line of work to overcome that limitation consists of identifying precipitation entities at more appropriate spatial scales, potentially at multi-scales. These entities could then substitute precipitation system as the experimental units for extracting attributes. In our view, there exist two potential techniques to address this problem. First, the extended watershed technique presented in [36] which provides a consistent and flexible framework for detecting convective storms over small spatial scales. Second, the multi-scale segmentation technique introduced in [37] which enables to split precipitation systems into sub-regions with specific textural properties.

Finally, there are a number of issues that we did not address in this study and that are left for future work. Firstly, the sensitivity of the results to the data length will be analyzed with the acquisition of new data or, if new data were not to become available, the application of resampling techniques such as bootstrap will be investigated. Secondly, this work aimed at classifying wind regimes at time t based on the weather conditions as seen by a weather radar at the same time t. It is planned to repeat the same study with lagged weather radar images, at time $t - k$, in order to examine the detection of early precipitation patterns. Thirdly, the temporal dimension of the sequence of images was not considered while each time series of precipitation attributes is characterized by a relatively strong autocorrelation. Further research will therefore be encouraged in this direction and data mining techniques dealing with autocorrelated data will receive specific attention.

ACKNOWLEDGMENT

This work was fully supported by the Danish Public Service Obligation (PSO) fund under the project “Radar@Sea” (contract PSO 2009-1-0226) which is gratefully acknowledged. Vattenfall is acknowledged for sharing the wind data from the Horns Rev 1 wind farm. The authors express their gratitude to the radar meteorologists from the Danish Meteorological Institute (DMI) for providing data from the Rømø radar and for their help.

REFERENCES

[1] C. Hasager, A. Peña, M. Christiansen, P. Astrup, M. Nielsen, F. Monaldo, D. Thompson, and P. Nielsen, "Remote sensing observation used in offshore wind energy," *IEEE J. Sel. Topics Appl. Earth Observ.*, vol. 1, pp. 67–79, 2008.

[2] P. Trombe, P. Pinson, T. Bøvith, N. Cutululis, C. Draxl, G. Giebel, A. Hahmann, N. Jensen, B. Jensen, N. Le, H. Madsen, L. Pedersen, A. Sommer, and C. Vincent, "Weather radars - The new eyes for offshore wind farms?" 2012, Working paper under review.

[3] L. Jones and C. Clark, "Wind integration - A survey of global views of grid operators," *In Proc. of the 10^{th} Int. Workshop on Large-Scale Integration of Wind Power into Power Systems, Aarhus, Denmark*, 2011.

[4] V. Akhmatov, C. Rasmussen, P. Eriksen, and J. Pedersen, "Technical aspects of status and expected future trends for wind power in Denmark," *Wind Energy*, vol. 10, pp. 31–49, 2007.

[5] P. Pinson, L. Christensen, H. Madsen, P. Sørensen, M. Donovan, and L. Jensen, "Regime-switching modelling of the fluctuations of offshore wind generation," *J. Wind Eng. Ind. Aerodyn.*, vol. 96, pp. 2327–2347, 2008.

[6] C. Vincent, G. Giebel, P. Pinson, and H. Madsen, "Resolving nonstationary spectral information in wind speed time series using the Hilbert-Huang transform," *J. Appl. Meteorol. Climatol.*, vol. 49, pp. 253–267, 2010.

[7] C. Vincent, P. Pinson, and G. Giebel, "Wind fluctuations over the North Sea," *Int. J. Climatol.*, vol. 31, pp. 1584–1595, 2011.

[8] H. Bluestein and M. Jain, "Formation of mesoscale lines of precipitation: Severe squall lines in Oklahoma during the spring," *J. Atmos. Sci.*, vol. 42, pp. 1711–1732, 1985.

[9] H. Bluestein, G. Marx, and M. Jain, "Formation of mesoscale lines of precipitation: Nonsevere squall lines in Oklahoma during the spring," *Mon. Weather Rev.*, vol. 115, pp. 2719–2727, 1987.

[10] M. Baldwin, J. Kain, and S. Lakshmivarahan, "Development of an automated classification procedure for rainfall systems," *Mon. Weather Rev.*, vol. 133, pp. 844–862, 2005.

[11] V. Lakshmanan and T. Smith, "Data mining storm attributes from spatial grids," *J. Atmos. Oceanic Technol.*, vol. 26, pp. 2353–2365, 2009.

[12] D. Gagne, A. McGovern, and J. Brotzge, "Classification of convective areas using decision trees," *J. Atmos. Oceanic Technol.*, vol. 26, pp. 1341–1353, 2009.

[13] G. Forney Jr, "The Viterbi algorithm," *Proc. of the IEEE*, vol. 61, pp. 268–278, 1973.

[14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer Series in Statistics, 2001.

[15] J. Kristoffersen, "The Horns Rev wind farm and the operational experience with the wind farm main controller," *In Proc. of the Offshore Wind Int. Conf. and Exhib., Copenhagen, Denmark*, 2005.

[16] R. Gill, S. Overgaard, and T. Bøvith, "The Danish weather radar network," *In Proc. of the 4^{th} European Conf. on Radar in Meteorol. and Hydrol., Barcelona, Spain*, 2006.

[17] V. Lakshmanan, J. Zhang, K. Hondl, and C. Langston, "A statistical approach to mitigating persistent clutter in radar reflectivity data," *IEEE J. Sel. Topics Appl. Earth Observ.*, 2012, available online.

[18] E. Cripps and W. Dunsmuir, "Modeling the variability of Sydney harbor wind measurements," *J. Appl. Meteorol.*, vol. 42, pp. 1131–1138, 2003.

[19] R. Davy, M. Woods, C. Russell, and P. Coppin, "Statistical downscaling of wind variability from meteorological fields," *Boundary-Lay. Meteorol.*, vol. 135, pp. 161–175, 2010.

[20] W. Zucchini and I. MacDonald, *Hidden Markov Models for time series: An introduction using R*. Chapman & Hall/CRC, 2009.

[21] P. Ailliot and V. Monbet, "Markov-Switching autoregressive models for wind time series," *Environ. Modell. & Softw.*, vol. 30, pp. 92–101, 2012.

[22] P. Bacher and H. Madsen, "Identifying suitable models for the heat dynamics of buildings," *Energy and Buildings*, vol. 43, pp. 1511–1522, 2011.

[23] T. Bøvith, "Detection of weather radar clutter," Ph.D. dissertation, Department of Informatics and Mathematical Modelling, Technical University of Denmark, Kgs. Lyngby, 2008, (ISBN: 87-643-0436-1).

[24] H. Wheater, V. Isham, C. Onof, R. Chandler, P. Northrop, P. Guiblin, S. Bate, D. Cox, and D. Koutsoyiannis, "Generation of spatially consistent rainfall data," Department of Statistical Science, University College London, Tech. Rep., 2000.

[25] J. Johnson, P. MacKeen, A. Witt, E. Mitchell, G. Stumpf, M. Eilts, and K. Thomas, "The storm cell identification and tracking algorithm: An enhanced WSR-88D algorithm," *Weather Forecast.*, vol. 13, pp. 263–276, 1998.

[26] I. Zawadzki, "Statistical properties of precipitation patterns," *J. Appl. Meteorol.*, vol. 12, pp. 459–472, 1973.

[27] E. Isaaks and R. Srivastava, *An Introduction to Applied Geostatistics*. Oxford University Press, 1989.

[28] E. Pebesma, "Multivariable geostatistics in S: the gstat package," *Comput. & Geosc.*, vol. 30, pp. 683–691, 2004.

[29] K. Suzuki, I. Horiba, and N. Sugie, "Linear-time connected-component labeling based on sequential local operations," *Comput. Visi. Image Und.*, vol. 89, pp. 1–23, 2003.

[30] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, pp. 679–698, 1986.

[31] A. Fitzgibbon, M. Pilu, and R. Fisher, "Direct least square fitting of ellipses," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, pp. 476–480, 1999.

[32] D. Sun, S. Roth, and M. Black, "Secrets of optical flow estimation and their principles," in *IEEE Conf. on Comput. Visi. and Pattern Recogn.*, 2010, pp. 2432–2439.

[33] M. Black and P. Anandan, "The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields," *Comput. Visi. Image Und.*, vol. 63, pp. 75–104, 1996.

[34] S. Li, *Markov Random Field modeling in image analysis*. Springer, 2009.

[35] H. Krolzig, *Markov-Switching Vector Autoregressions: modelling, statistical inference, and application to business cycle analysis*. Springer, 1997.

[36] V. Lakshmanan, K. Hondl, and R. Rabin, "An efficient, general-purpose technique for identifying storm cells in geospatial images," *J. Atmos. Oceanic Technol.*, vol. 26, pp. 523–537, 2009.

[37] V. Lakshmanan, R. Rabin, and V. DeBrunner, "Multiscale storm identification and forecast," *Atmos. Res.*, vol. 67, pp. 367–380, 2003.

Pierre-Julien Trombe received the M.Sc. degree in applied mathematics from the National Institute for Applied Sciences (INSA), Toulouse, France, in 2005. He is currently pursuing the Ph.D. degree at the Technical University of Denmark, Kgs. Lyngby, Denmark.

His research interests include among others forecasting, image analysis and renewable energies.

Pierre Pinson (M'11) received the M.Sc. degree in applied mathematics from the National Institute for Applied Sciences (INSA), Toulouse, France, and the Ph.D. degree in Energy from Ecole des Mines de Paris, Paris, France.

He is an Associate Professor in stochastic energy systems at the Informatics and Mathematical Modeling Department of the Technical University of Denmark, Kgs. Lyngby, Denmark. His research interests include among others forecasting, uncertainty estimation, optimization under uncertainty, decision sciences, and renewable energies.

Henrik Madsen received the Ph.D. degree in statistics at the Technical University of Denmark, Kgs. Lyngby, Denmark, in 1986.

He was appointed Professor in mathematical statistics in 1999. He is an elected member of the International Statistical Institute (ISI), and he has participated in the development of several ISO and CEN standards. His research interests includes analysis and modeling of stochastic dynamics systems, signal processing, time series analysis, identification, estimation, grey-box modeling, prediction, optimization, and control, with applications mostly related to energy systems, informatics, environmental systems, bioinformatics, biostatistics, process modeling, and finance.