

Turbulent coherent structures in the atmospheric surface layer: detection on Doppler lidar observations by supervised machine learning

Perrine Maynard^(a,*), Elsa Dieudonné^(a), Anton Sokolov^(a), Hervé Delbarre^(a)

^(a) *Laboratory of Physico-Chemistry of the Atmosphere (LPCA), UR 4493, University of Littoral Opal Coast (ULCO), Dunkirk, France.*

^(*) Perrine.maynard@univ-littoral.fr

Abstract: Turbulent structures, particularly coherent ones like streaks, significantly influence turbulent fluxes and the dispersion of pollutants in the surface layer. These structures can be observed directly on the horizontal scans of a Doppler lidar, performed during a 13-month campaign in Dunkirk, France, an industrial city on the North Sea shore. About forty thousand quasi-horizontal scans were recorded, capturing two main types of coherent structures: organized and disorganized streaks (a third category named "others" accounting for the absence of structures). An automated classification method was developed for classifying the large dataset of lidar images. The images were pre-processed, notably for retrieving the turbulent part of the wind [1]. Each image was represented by a vector of features designed to highlight the streak patterns and computed in 3 steps: (1) Gray-level Co-occurrence Matrices, (2) texture parameters such as contrast, homogeneity, correlation, or energy [2], and (3) "curve parameters" [1]. A training set consisting of four hundred scans was built to train supervised learning classification algorithms. The Quadratic Discriminant Analysis classifier has the best performance among the four classification algorithms proved; it classified the training set with a cross-validation error of only 5.2 %. All four classifiers successfully discriminated the categories, with about 61% of the scans classified as coherent structures, including about 31% of organized streaks and 30% of disorganized streaks.

1. Introduction

Exploring turbulence within the surface layer is crucial for enhancing our grasp of atmospheric flow dynamics. This helps improving how surface-atmosphere interactions are represented in numerical weather forecasting and chemistry-transport models, thereby enhancing the precision of weather and air quality predictions. A detailed comprehension of the small-scale dynamics is also essential for tackling environmental issues such as urban heat islands. Precise predictions and analyses of turbulence could aid in designing more efficient wind farms, improving their performance forecast, and evaluating their effects on surface layer turbulence. Turbulent streaks, convective rolls, and gravitational waves are the most frequently examined types of coherent structures [3]. Streaks, which are formed due to wind shear, manifest as alternating parallel bands of faster and slower wind in the surface layer [4]. These bands are

often associated with vortices or semi-linear swirls that generally align with the prevailing wind direction, with adjacent vortices rotating in opposite directions [4].

This study aims to explain how coherent structures like streaks in the surface layer are influenced by surface conditions and meteorological factors, utilizing a substantial dataset that allows for the computation of annual statistics. The observations consisted of almost 40,000 quasi-horizontal scans from a Doppler lidar that operated on the North Sea shore for 13 months. The initial phase of the study involved classifying the dataset through the supervised machine learning (SML) techniques introduced by Cheliotis et al. [1] with texture parameters.

2. Experimental set-up and data pre-processing

This study took place in Dunkirk (France), a coastal site located between the English

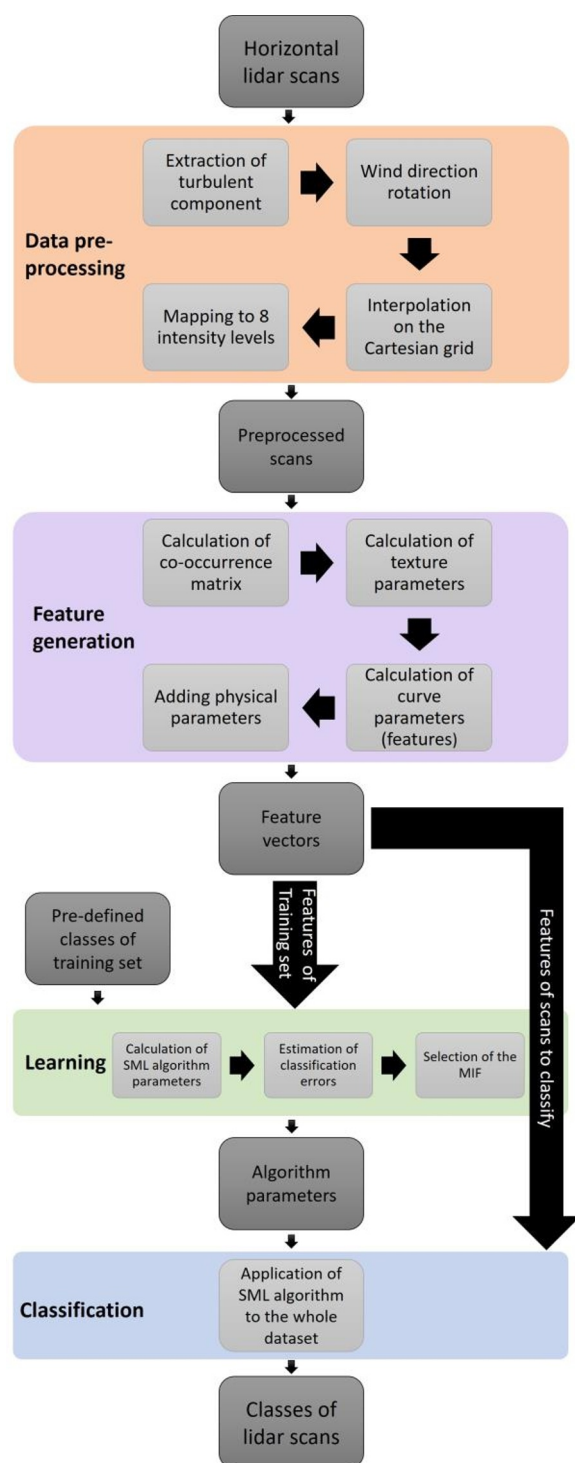


Figure 1. Phases and steps of the data processing flow.

Channel and the open North Sea. The regional topography is flat with an average elevation of 5 m above sea level. The area covered by the lidar scans includes industrial sites and urban areas.

The scanning Doppler lidar used in this research was a WindCube Scan WCS100 from Leosphere / Vaisala. It was installed on the roof

of a building, 1 km from the coastline (51°02'15.3"N 2°21'57.6" E), 16 m above ground level, and 23 m above sea level. This lidar operates at a wavelength of 1.543 μm and is designed for radial wind speeds ranging from -30 to $+30$ $\text{m}\cdot\text{s}^{-1}$ with an accuracy of ± 0.1 $\text{m}\cdot\text{s}^{-1}$. The observations used in this study were almost horizontal conical sweeps of the Plan-Position Indicator (PPI) type, recorded with an elevation of 1° , an azimuthal resolution of 2° , and an axial resolution of 50 m. A blind zone of 100-m radius existed at the scan center, and the maximal range was 7.2 km. Continuous monitoring occurred from May 18, 2021, to June 13, 2022, giving 37,809 PPI scans. With an accumulation time of 1 s per beam, each scan lasted for 180 seconds, and the system repeated its measurement cycle every 14 minutes.

The data processing flow was divided into four phases (Fig.1) as in Cheliotis et al. [1]: (i) the raw lidar data was converted into images, (ii) statistical parameters were computed from each image, (iii) the SML algorithm was trained and (iv) the actual classification was performed (Sec. 3 for iii and iv).

The preparation of the lidar images involved four steps. Firstly, the turbulent component was isolated from the radial wind measured by the lidar using the VAD method [1]. Secondly, the images were standardized by rotating them to align the mean wind direction with the y-axis, thus simulating a wind always blowing from the North. Thirdly, the data points were interpolated from a polar to a Cartesian grid to get square pixels. And fourthly, the turbulent radial wind speed values were cropped into the ± 1 $\text{m}\cdot\text{s}^{-1}$ interval and mapped into a uniform 8-level scale, to amplify the textural characteristics.

The second phase of the data processing was designed to extract the relevant information from the images and their texture parameters while reducing the amount of data, which is done by computing statistical parameters, called features, from the images. Indeed, supplying the classification process with irrelevant and redundant data may result in overfitting (when a machine learning model conforms too tightly to the training data, it ends up absorbing noise and irregularities instead of the patterns). It is furthermore inefficient in terms of computation time.

The transformation of the 8-level images into numerical descriptors entailed three stages (Fig.1). Firstly, grey-level co-occurrence matrices established by Haralick et al. [2] were computed from the image for all pixel-pair geometric configurations up to the 50th neighbor. Secondly, texture parameters were derived from these matrices, namely contrast, homogeneity, correlation, and energy, as in Haralick et al. [2]. Thirdly, curve parameters were derived from the texture parameters, like Cheliotis et al. [1]. A few meteorological parameters were added (e.g. the mean wind speed) and each 393×393-pixel image was finally characterized by a 604-element feature vector.

Upon visually inspecting the dataset, two distinct categories of streaks emerged a classification unseen in the literature, but previous observational studies about streaks are rare and limited to a few days at most e.g. [5]. The term 'disorganized streaks' denoted elongated formations that exhibited not linearly and parallel, appearing more chaotic or mingled with incoherent areas (Fig. 2a). In contrast, 'organized streaks' was used to describe structures that were elongated, highly linear, and parallel (Fig. 2b). All other types of formations that did not fit into these categories were collectively assigned to a group called 'others' (Fig. 2c).

A manually classified training set consisting of 399 scans (~1% of the whole dataset) was built to train the supervised learning classification algorithm. It is essential to ensure that the training set accurately represents the whole dataset, or the training results may not be transposable to the whole dataset. A great care was thus paid to select cases as evenly distributed as possible in terms of season and time of the day, with proportions matching the estimated shares of the different categories in the whole dataset.

3. Turbulent patterns classification using supervised machine learning (SML)

If the features have been well designed, the different classes of the training ensemble should correspond to well-separated clusters of points along some of the 604 dimensions of the feature space. At the learning step, the role of the SML algorithm parameters is to identify the small

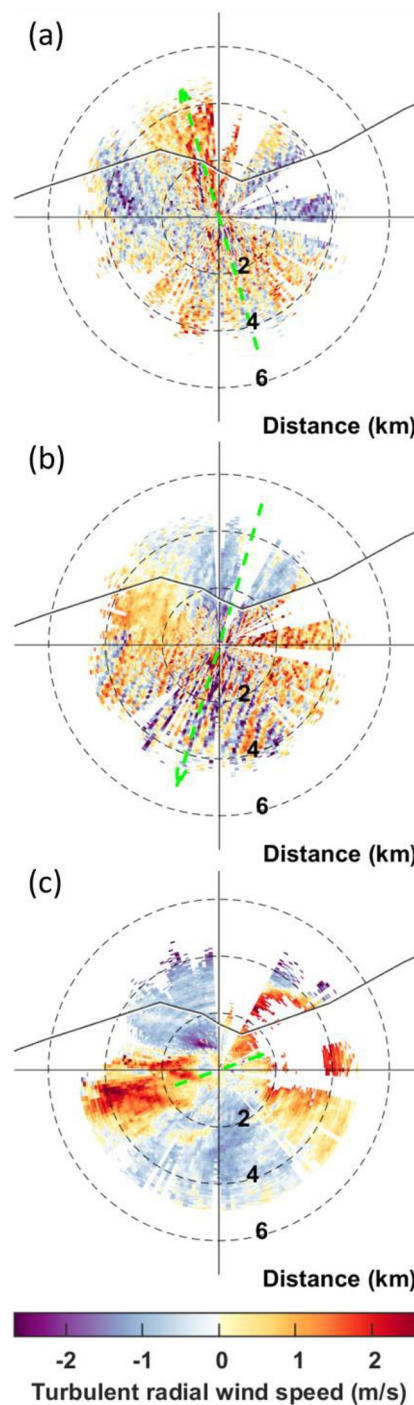


Figure 2. Example scans showing the different classes of lidar images: (a) disorganized streaks, (b) organized streaks, and (c) absence of coherent structures ('others'). The solid black line features the coastline; the green arrow shows the mean wind.

subset of features, called most informative features (MIF), that allows to best separate the classes and thus minimize the classification error on the training set. At the classification

step, the algorithm is then applied to classify the unknown data.

In this study, different SML algorithms were evaluated: Quadratic Discriminant Analysis (QDA), Error-Correcting Output Codes Support Vector Machine (ECOC-SVM), k-nearest Neighbors (KNN), and Artificial Neural Networks (ANN). The classification error was assessed using the 10-fold cross-validation method, which consists of training the algorithm using only 90 % of the training set, then checking the accuracy of the classification on the remaining 10 %, and repeating the process 10 times until all the training sets have been classified.

The lowest classification error of 5.2 % was obtained for the QDA algorithm, utilizing nine features as MIF. The ECOC-SVM algorithm reached an error of 7.0 % with fifteen MIF. For the KNN method with 9 neighbors (K=9), the error registered at 9.1 % when thirteen MIF were used, and the ANN method recorded an error of 9.6 % with fifteen MIF. QDA was therefore selected for classifying the whole dataset. The ‘others’ cases were almost all correctly classified; the error mainly resulted from inversions between the organized and disorganized streaks categories. Although these two categories are quite similar looking, the algorithm still largely succeeded in differentiating them.

In the whole dataset, 61 % of the scans were identified as streaks, the organized ones corresponding to 31 % and the disorganized ones to 30 % (Fig. 3). The remaining 36 % of the scans fell into the ‘others’ category, with the final 3 % being unclassifiable due to insufficient data to compute the features.

4. Conclusion and Perspectives

The main goal of this study is to understand how weather conditions affect turbulent streaks in the surface layer. The first phase was to detect automatically the two identified types of streaks – organized and disorganized – on the large dataset of lidar images obtained from quasi-horizontal scans recorded in the surface layer. A classification method based on SML algorithms was developed that required two stages of pre-processing of the dataset, to derive statistical parameters (features) representing the images. This classification method was effective despite the very different weather

conditions existing during the 13 months covered by the database.

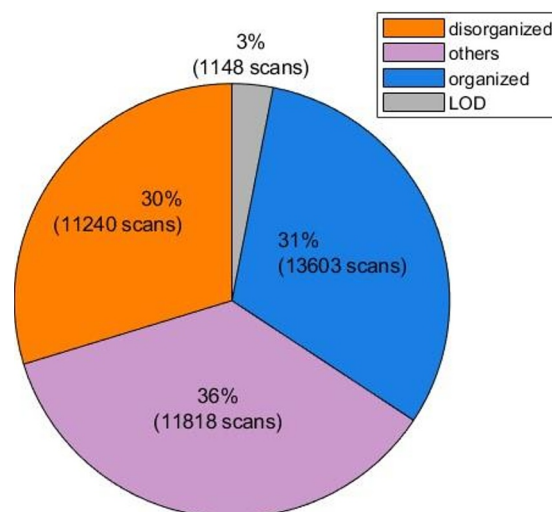


Figure 3. Results of the classification on the whole dataset. ‘LOD’ (lack of data) means the amount of valid data was insufficient to compute the statistical parameters and thus, classify the image.

The current error rate is suitable for the next research stage which is linking the streak patterns with environmental factors (wind shear, atmospheric stability...) and calculating their dimensions relative to the boundary layer depth, to study their development.

5. References

- [1] Cheliotis, I., E. Dieudonné, H. Delbarre, A. Sokolov, E. Dmitriev, P. Augustin and M. Fourmentin, “Detecting Turbulent Structures on Single Doppler Lidar Large Datasets: An Automated Classification Method for Horizontal Scans,” *Atmospheric Measurement Techniques*, 2020.
- [2] Haralick, R. M., K. Shanmugam, and I. Dinstein, “Textural Features for Image Classification,” *IEEE Transactions on Systems, Man, and Cybernetics*, 1973.
- [3] Young, G. S., Kristovich, D. A. R., Hjelmfelt, M. R., and Foster, R. C.: Supplement to Rolls, Streets, Waves, and More, *Bulletin of the American Meteorological Society*, 2002.
- [4] Khanna, S. and Brasseur, J. G.: Three-Dimensional Buoyancy- and Shear-Induced Local Structure of the Atmospheric Boundary Layer, *Journal of the Atmospheric Sciences*, 1998.
- [5] Träumner, K., Damian, Th., Stawiarski, Ch., and Wieser, A.: *Turbulent Structures and Coherence in the Atmospheric Surface Layer, Boundary-Layer Meteorology*, 2015.