Automated Classification of Regional Meteorological Events in a Coastal Area Using In Situ Measurements

Anton Sokolov, Egor Dmitriev, Cyril Gengembre, Hervé Delbarre

Automated Classification of Regional Meteorological Events in a Coastal Area Using In Situ Measurements

ANTON SOKOLOV

Laboratory for Physico-Chemistry of the Atmosphere, University of Littoral Cote d'Opale, Dunkirk, France

EGOR DMITRIEV

Institute of Numerical Mathematics, Russian Academy of Sciences, and Moscow Institute for Physics and Technology, Moscow, Russia

CYRIL GENGEMBRE AND HERVÉ DELBARRE

Laboratory for Physico-Chemistry of the Atmosphere, University of Littoral Cote d'Opale, Dunkirk, France

(Manuscript received 19 July 2019, in final form 11 February 2020)

ABSTRACT

The problem is considered of atmospheric meteorological events' classification, such as sea breezes, fogs, and high winds, in coastal areas. In situ wind, temperature, humidity, pressure, radiance, and turbulence meteorological measurements are used as predictors. Local atmospheric events of 2013–14 were analyzed and classified manually using data of the measurement campaign in the coastal area of the English Channel in Dunkirk, France. The results of that categorization allowed the training of a few supervised classification algorithms using the data of an ultrasonic anemometer as predictors. The comparison was carried out for the *K*-nearest-neighbors classifier, support vector machine, and two Bayesian classifiers—quadratic discriminant analysis and Parzen–Rozenblatt window. The analysis showed that the *K*-nearest-neighbors and quadratic discriminant analysis classifiers reveal the best classification accuracy (up to 80% correctly classified meteorological events). The latter classifier has higher calculation speed and is less sensitive to unbalanced data and the overtraining problem. The most informative atmospheric parameters for events recognition were revealed for each algorithm. The results obtained showed that supervised classification algorithms contribute to automation of processing and analyzing of local meteorological measurements.

1. Introduction

Information about local meteorological phenomena is widely used in making managerial decisions related to regional development. One of the most challenging domains for atmospheric study are coastal zones, areas with complex orography and industrial areas located near large settlements. Mesoscale circulation in the boundary layer of the atmosphere has a significant effect on the transport and dispersion of pollutants. In particular, the development of breeze circulation contributes to the accumulation of contaminants in the area up to 20–30 km inland. The appearance of fog leads to impairing visibility and to the formation of smog. Multiple dangerous situations can arise from a sudden increase of the wind speed (like squalls and local storms).

Dangerous meteorological phenomena occurring at regional scales often have poor predictability at synoptic time scales. Usually, the probabilities of weather samples in selected regions are estimated based on in situ and remote measurements or by modeling. Local circulation patterns, like breezes, should be taken into account in this type of analysis. The analysis of the sea-breeze impact on the concentration and dispersion of different atmospheric pollutants presented in Mavrakou et al. (2012) revealed the intensification of dispersion processes during the sea-breeze days. Considering mesoscale circulation under different synoptic conditions allows a better understanding of physical mechanisms of atmospheric pollutants' transfer and dispersion in densely populated industrial coastal areas (Boyouk et al. 2011; Sokolov et al. 2016; Kambezidis et al. 1998).

Corresponding author: Anton Sokolov, anton.sokolov@univ-littoral.fr

DOI: 10.1175/JTECH-D-19-0120.1

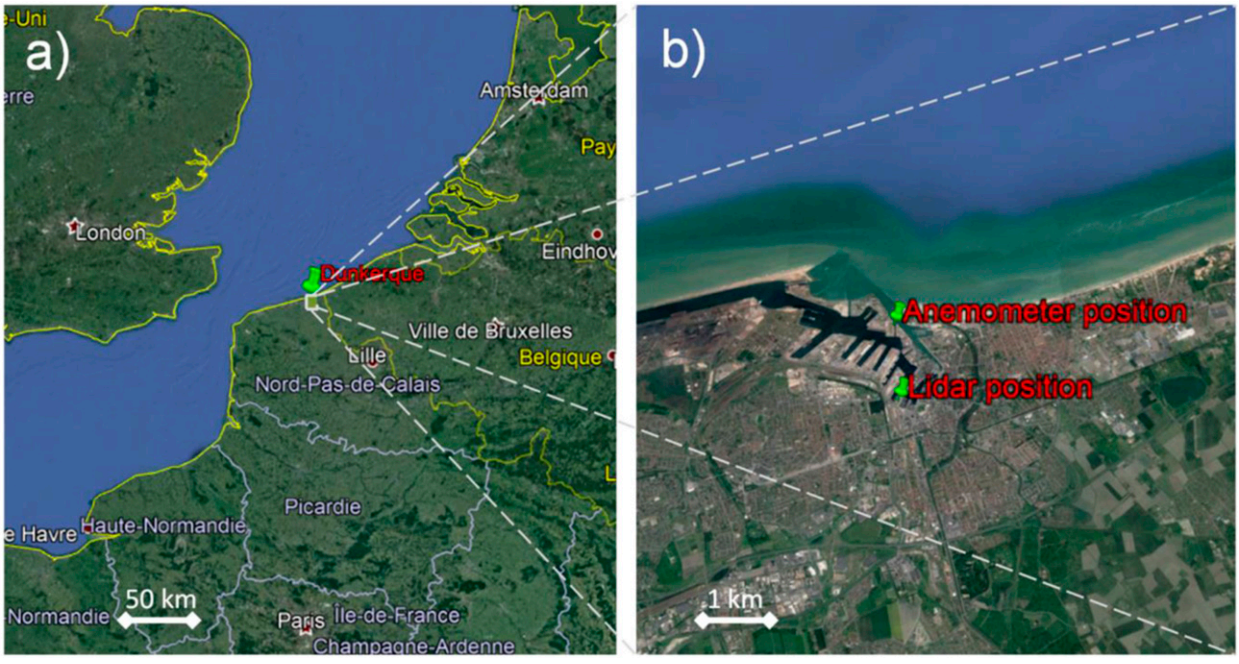

FIG. 1. Position of (a) Dunkirk region and (b) the measuring instruments, plotted with Google Earth Pro (map data: copyright 2020 GeoBasis-DE/BKG, copyright 2020 Google; data are from SIO, NOAA, U.S. Navy, NGA, and GEBCO; image is from Landsat/Copernicus).

Sometimes the detection of local meteorological phenomena is a complicated scientific problem demanding the expert analysis of multiple in situ, remote and atmospheric modeling data. Thus, an important work should be done to have a reliable dataset of classified meteorological events suitable for statistical study. This classification task becomes even more complicated for the statistical analysis of large meteorological databases. In this article, we suggest to simplify and automate this kind of analysis using machine learning (ML) techniques, namely supervised classification.

The main steps of the proposed algorithm and corresponding sections are the following:

1) Preparing of meteorological data. The dataset is described in section 2.
2) Classification of meteorological events by an expert. The information about the expert classification could be found in section 3.
3) Selection of most informative predictors (or features) minimizing the classification error for each ML algorithm for the restricted and for the whole dataset. Theoretical backgrounds and some aspects of ML methods are presented in section 4. The procedure of retrieving the optimal sequence of features could be found in section 5.
4) Evaluation of performance of selected supervised ML algorithms by estimation of the total classification error, other classification characteristics and corresponding confidence intervals. The theoretical information on classification characteristics and bootstrapping technique presented in section 5.

The classification result of different ML algorithms with restricted and whole datasets could be found in section 6.

2. Domain of study and measurements

The present study uses data collected in a field campaign performed in Dunkirk (see Fig. 1a), situated in the north of France on the coastal area of the English Channel. This area is highly urbanized and industrialized, and suffers from local pollution sources such as steel industry, oil refining, sea and land transport emissions. It is also influenced by the pollution of Lille and Paris in France, London (England), and Rotterdam (the Netherlands) agglomerations.

The region is characterized by nearly flat orography, and by strong atmospheric perturbations on the border of land–sea–atmosphere. These complex atmospheric processes are responsible for transport and dispersion of air pollution and govern local air quality. Multiple studies were conducted for the region on atmospheric dynamics and pollution dispersion (Talbot et al. 2007; Boyouk et al. 2011; Xiang et al. 2012).

We used the data of the measuring campaign that lasted from July 2013 until August 2014. To characterize

TABLE 1. Main meteorological and micrometeorological parameters measured by ultrasonic anemometer (rows 1–8) and the weather station (rows 9–13).

	Parameter	Description	Units
1	HorWind	Horizontal wind speed	$m\,s^{-1}$
2	DirWind	Wind direction with respect to true north, from which the wind is coming (0° corresponds to north, 90° to east, 180° to south, and 270° to west)	°
3	UWind	Wind component from the west to the east	$m\,s^{-1}$
4	VWind	Wind component from the south to the north	$m\,s^{-1}$
5	Ustar	Friction or shear velocity	$m\,s^{-1}$
6	SigW	Vertical wind speed fluctuations (RMS of w)	$m\,s^{-1}$
7	HF	Vertical heat flux	$°C{\cdot}m\,s^{-1}$
8	MF	Vertical momentum flux	$m^2\,s^{-2}$
9	T	Temperature	°C
10	RH	Relative humidity	%
11	P	Pressure	10^2 Pa
12	R	Solar radiation	$W\,m^{-2}$
13	RF	Rainfall	10^{-3} m

the local state of the atmosphere and phenomena occurring on a microscale, an instrument should be employed capable of measuring multiple atmospheric parameters at a fine time scale of less than one hour. Therefore, we chose the ultrasonic anemometer coupled with a weather station (see Fig. 1b) able to perform micrometeorological measurements in the surface layer with a time resolution of 15 min (see Fig. 1 for the position of the anemometer). We used 3D sensor USA-1 of Metek (see http://metek.de/product/usonic-3-scientific/), which can measure a few turbulence parameters in addition to the horizontal wind. All atmospheric parameters measured in situ are presented in Table 1.

Wind measurements obtained by the Doppler lidar profiler are often indispensable in addition to in situ measurements to analyze the processes related to the dynamics of the lower troposphere. We used WINDCUBE 100S model produced by the Leosphere company (http://www.leosphere.com/products/3d-scanning/windcube-100s200s400s-turbulence-wind-lidar) (see Fig. 1 for the position of lidar). The lidar resolution is 50 m; the range depends on atmospheric state and allows having a dense data below 500 m. Both plan position indicator (PPI) and the range–height indicator (RHI) scanning patterns were acquired.

In an anticyclonic situation, sunshine is essential for the development of thermals and the formation of the atmospheric boundary layer. It is also responsible for the occurrence of local weather phenomena such as the sea-breeze circulation. In this context, we applied a model of solar radiation (see Benkaciali and Gairaa 2014; Iqbal 1983; Sen 2008) to analyze the evolution of solar radiation during the different meteorological events and compared the expected radiation in the case of a clear and cloudless sky (modeling), with the experimental data.

3. Expert classification of mesoscale meteorological events

A local atmospheric state was analyzed manually using the 1-yr dataset of ultrasound anemometer, weather station and Doppler lidar wind profiler. Recognition criteria for detection of meteorological events were established and events were identified. Here we describe briefly the procedure of manual events classification by an expert [see Gengembre (2018) for details]. In some cases, the Infoclimat meteorological site (https://www.infoclimat.fr/) was used to analyze the global atmospheric situation.

a. Sea-breezes detection

As it was shown by few studies carried out in Dunkirk (Boyouk et al. 2011; Talbot et al. 2007), the breeze circulation is associated with sunny weather, low cloudiness and low wind. The sea-breeze start was detected under anticyclonic meteorological conditions (Miller et al. 2003; Simpson 1994) by a significant transition in meteorological parameters such as wind direction and speed, temperature, relative humidity and solar radiation. Previous studies have also shown that it is difficult to identify the end of a sea breeze.

However, it is sometimes difficult to recognize breezes using only ground measurements. In those ambiguous situations the wind profiles obtained by Doppler lidar allows detecting breezes. Using lidar data, we could clearly see patterns related to the development of thermals and increasing of the atmospheric boundary layer before the breeze start, then we observe an important transition of the profile of direction of the synoptic wind and finally the breeze gravity current established until about 300 m above sea level.

To summarize, the method of the sea-breeze detection during the long campaign consists in selecting,

at first, the anticyclonic and sea-wind periods (from northwest to northeast). Then the measured solar radiation is compared to the model in order to choose the low cloud periods. Then, the periods of rapid transitions of meteorological parameters (temperature, relative humidity, and wind) were analyzed and breeze circulation was validated using vertical wind profiles.

b. *Detection of fog events*

During an episode of fog, the range of the Doppler lidar is decreased. The event recognition method is then restricted to analyzing only the ground data of the ultrasonic anemometer and the weather station.

Conventionally, visibility is the parameter used to detect fog. This type of measurement was not available in our study, but investigations have shown that the relative humidity threshold is a relevant parameter for identifying a fog, although no precise threshold exists. With regard to the relative humidity indexed in the literature (Doyle and Dorling 2002; Ding and Liu 2014; Liu et al. 2012), we decided to apply as a first filtering criterion a relative humidity threshold.

Like for the sea breeze, previous work has shown marked transitions in meteorological parameters during the setup of fog. Relative humidity increases to the near-saturation point and the difference between air temperature and dewpoint becomes small during the event. Goswami and Sarkar (2015) showed this difference is as low as 2 K. A fog is also characterized by a low wind speed, which is generally less than 2 $m \cdot s^{-1}$ (Dupont et al. 2012; Ye et al. 2015; Wang et al. 2015; Degefie et al. 2015). The Doppler lidar, which does not measure wind profiles during the fog episodes, could nevertheless provide useful information by decreasing of measurements range, or even by the absence of a signal.

The method for detecting fog events was therefore first selecting the periods with the relative humidity is greater than 85%. Then we analyzed the transitions of the meteorological parameters (temperature, dewpoint, and wind).

c. *Detection of high wind events*

To detect high-wind events (HWE), the lower wind speed threshold of 14 $m \cdot s^{-1}$ was introduced. Thus, HWE corresponds to the seventh–tenth categories of the Beaufort scale (see Saucier 1955), associated with remote low pressure centers. The upper threshold corresponding to the "storm" domain was set at 24.5 $m \cdot s^{-1}$. Relying on the data obtained during the local measurement campaign, we have found that the horizontal wind speed measured by the ultrasonic anemometer approached the upper threshold but never exceeded it.

The lidar measurements reveal an acceleration of the wind speed with altitude during HWEs even though the lidar range could be reduced due to the presence of clouds. Negative vertical velocity detected by lidar could be related to the phenomenon of "precipitous descents" (see Risi et al. 2016; Kessler 1995).

To summarize, studied meteorological phenomena were identified by values and transitions of meteorological parameters measured locally by the ultrasonic anemometer, the weather station and the lidar profiler. By implementing those manual weather recognition methods, the following events were identified during the long campaign in Dunkirk:

- 36 sea breezes with an average duration of 6.5 h,
- 15 events of fogs spread over 20 days with an average duration of 9 h, and
- 12 HWEs with an average duration of 5 h.

4. Basic supervised classification methods

The supervised classification problem can be formulated as follows. Let $\mathbf{X}$ be the set of features (M-dimensional vectors) and $\mathbf{Y}$ be the finite set of Q object labels. In our case, features are atmospheric measurements, and labels are types of meteorological events. We need to construct an algorithm $s = a(\mathbf{x})$, where $\mathbf{x} \in \mathbf{X}$ and $s \in \mathbf{Y}$. Unknown parameters of the algorithm are estimated using prior information represented as the finite set

$$\mathbf{X}^N = \{\mathbf{x}_i, y_i\}_{i \in [\![1,N]\!]}$$

(training set) consisting of pairs of elements from sets $\mathbf{X}$ and $\mathbf{Y}$, with N being the number of such pairs. Thus, the algorithm $a(\mathbf{x})$ should be optimal in some certain sense on $\mathbf{X}^N$. We applied several basic classification algorithms of different complexity: K nearest neighbors (KNN), error-correcting output codes with support vector machine (SVM), and two Bayesian classifiers (BC): quadratic discriminant analysis (QDA) and the Parzen–Rosenblatt window method (PRW).

a. *K nearest neighbors*

Similarity-based classification algorithms imply the calculation of certain similarity function characterizing distances between objects. KNN method is the simplest classifier of such a type. In accordance with this method, the object to be classified is associated with the class owning the most number of objects similar to it.

Let us sort the training set in the order of ascending distance $\rho(\mathbf{u}, \mathbf{x}_i)$ between the classified feature $\mathbf{u} \in \mathbf{X}$ and training features $\mathbf{x}_i \in \mathbf{X}$; that is, $\rho(\mathbf{u}, \mathbf{x}_1) \leq \cdots \leq \rho(\mathbf{u}, \mathbf{x}_N)$. The algorithm of K nearest neighbors can be represented by the following formula:

$$a(\mathbf{u}) = \arg\max_{y \in \mathbf{Y}} \sum_{i=1}^{N} [y_i = y][i \leq K],$$

where $a(\mathbf{u})$ is the resulting class, square brackets signify the indicator function under the statement: 1 for correct statement and 0 for incorrect one, y_i are object labels from the sorted training set, and K is the number of neighbors used. Thus, the classified feature vector $\mathbf{u}$ is associated with the most abundant class among K of its nearest neighbors.

For small values of K, the discriminant surface (see Dreyfus 2005) has as a rule a complex shape and can be multiply connected that allows solving sophisticated classification problems. On the other hand, this surface is very sensitive to small changes in the training sample, and accordingly the result of classification will be unstable. For large values of K, on the contrary, the algorithm is excessively stable and degenerates into a constant. Thus, the extreme values of K are undesirable. The optimal value K can be determined by using the leave-one-out cross-validation method.

The classification process described above is computationally expensive because it implies the explicit storage of all training objects and an exhaustive search on them. This problem becomes especially important in the case of a large number of training samples used. In this paper, we employed a more effective search of nearest neighbors by using the effective data structure (see Friedman et al. 1977).

b. *Error-correcting output codes with support vector machine*

Among the known approaches used to solve the supervised classification problems, there is a group of methods that were originally designed to divide the processed data into two groups (binary classification). One of the known effective methods of this kind is SVM. The method of error-correcting output codes (ECOC; see Dietterich and Bakiri 1995) is based on several approaches from the information coding theory and allows extending binary classifiers to the multiclass case. The multiclass classifier based on the combination of ECOC and SVM methods can be described as follows.

Let us introduce the coding design matrix $\mathbf{C} = (c_{ij})$ of size $Q \times L$ with elements 1, −1, and 0. The lines of this matrix are unique codes of Q considered classes. The L columns of matrix $\mathbf{C}$ define a series of different binary learners deciding between two composite classes constructed from the initial ones. Thus, for each column, the first composite class aggregates initial classes corresponding to 1, the composite class aggregates classes corresponding to −1, and classes not participating in the binary classification correspond to 0.

The coding stage of the ECOC classifier consists in the consecutive application of the mentioned L learners to the query feature sample $\mathbf{x}$. Thereby we obtain the code of some unknown class corresponding to $\mathbf{x}$. On the decoding stage, this code is compared with the codes of initial Q classes by using some selected measure of distance, for example, Euclidian, Chebyshev, or Hamming. Thus, the query sample $\mathbf{x}$ is assigned to the class label with the most similar code.

The choice of the coding-design matrix significantly affects the accuracy and calculation speed of the ECOC classifier. In this paper we have used the one-versus-one coding design, which can be defined as

$$\mathbf{C} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 0 & -1 & -1 \end{pmatrix}$$

for the case of $Q = 4$ initial classes. This coding design provides the balance between the accuracy and calculation efficiency (Dmitriev et al. 2018). The Hamming distance is chosen as the measure of similarity between the codes of classes.

The kernel soft-margin SVM (see Scholkopf and Smola 2002) is used in the ECOC algorithm as the basic binary classifier. The linear soft-margin SVM allows finding the most distant parallel hyperplanes $(\mathbf{w}, \mathbf{x}) - w_0 = 1$ and $(\mathbf{w}, \mathbf{x}) - w_0 = -1$ in the multidimensional feature space ($\mathbf{w}$ and w_0 are parameters of the hyperplanes) separating the given pair of classes labeled as 1 and −1. The hyperplanes pass through the area containing boundary points (support vectors) of feature distributions of these classes. Optimization is performed by taking into account penalties for misclassification of the boundary points. SVM is easily generalized to the case of nonlinear separating surfaces by replacement of the scalar product $(\mathbf{x}', \mathbf{x}'')$ by a kernel $K(\mathbf{x}', \mathbf{x}'')$. In this case, the classification algorithm has the form

$$a(\mathbf{x}) = \operatorname{sign}\left[\sum_{i=1}^{N} \lambda_i(\mathbf{w}) y_i K(\mathbf{x}_i, \mathbf{x}) - w_0\right].$$

Training of this algorithm consists in the estimation of the parameters λ (and associated $\mathbf{w}$) and w_0. The Gaussian kernel

$$K(\mathbf{x}', \mathbf{x}'') = \exp\left(-\frac{\|\mathbf{x}' - \mathbf{x}''\|^2}{2\sigma^2}\right)$$

is used in this paper (σ is the kernel scale). This kernel is well suited for outlining classes having a complex shape in the feature space at the acceptable calculation speed.

In contrast with labels assigned to samples $\mathbf{x}$, the reliability of classification results is not the same and

depends on the position of $\mathbf{x}$ in the feature space. Thus classification score $s(\mathbf{x})$ is frequently considered as the extended output of the classification algorithm. For SVM classifier, the function $s(\mathbf{x})$ indicates the normalized distance from the sample $\mathbf{x}$ to the discriminant surface in the area of relevant class. Let us also introduce the function

$$g(y_B, \mathbf{x}) = \max\frac{\{0, 1 - y_B s(\mathbf{x})\}}{2},$$

where $y_B \in \{1, -1\}$ is a label in the binary classification problem. This function characterizes the loss when classifying the sample $\mathbf{x}$ (the binary loss). Thus for $s(\mathbf{x}) \geq 1$ we have zero binary loss. For the correctly classified samples, the binary loss does not exceed 0.5.

The classification algorithm described above can be formulated in the more general form (Escalera et al. 2010). On the coding stage, we calculate the classification scores $s_j(\mathbf{x})$ for each of the L binary learners defined by coding-design matrix $\mathbf{C}$ and binary losses $g[c_{ij}, s_j(\mathbf{x})]$ for each of the classes y_i considered in the initial multiclass problem. The decoding stage can be expressed by the formula

$$a(\mathbf{x}) = \arg\min_{y_i} \frac{\sum_{j=1}^{L} |c_{ij}| g\{[c_{ij}, s_j(\mathbf{x})]\}}{\sum_{j=1}^{L} |c_{ij}|},$$

where the minimum search is performed on the class index i and the class label y_i is the output. Thus, the algorithm selects the class corresponding to the minimum average loss.

c. *Bayesian classification, QDA and PRW*

The use of BC implies that features can be regarded as random variables. In this paper, we consider the continuous distribution of features. The general form of the BC algorithm is

$$a(\mathbf{x}) = \arg\max_{\mathbf{y}\in\mathbf{Y}} P_{\mathbf{y}} p_{\mathbf{y}}(\mathbf{x}),$$

where $P_{\mathbf{y}}$ is the prior probability of class y and $p_{\mathbf{y}}(\mathbf{x})$ is the probability density function (PDF) of features of this class. In the case in which $P_{\mathbf{y}}$ and $p_{\mathbf{y}}(\mathbf{x})$ are known exactly, the general form of BC is optimal because the solution obtained has the minimum total probability of the classification error. The training of BC consists in the estimation of distributions of features (and optionally prior probabilities) for all considered classes using the data from the training set.

QDA (or normal BC) (Hastie et al. 2008) is the parametric approach implying that PDFs belong to the family of normal distributions $p_{\mathbf{y}}(\mathbf{x}) \in N(\boldsymbol{\mu}_{\mathbf{y}}, \boldsymbol{\Sigma}_{\mathbf{y}})$, where $\boldsymbol{\mu}_{\mathbf{y}}$ is the expectation vector and $\boldsymbol{\Sigma}_{\mathbf{y}}$ is the covariance matrix of features of the class y. Estimates $\hat{\boldsymbol{\mu}}_{\mathbf{y}}$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{y}}$ of the parameters $\boldsymbol{\mu}_{\mathbf{y}}$ and $\boldsymbol{\Sigma}_{\mathbf{y}}$ can be obtained from the principle of maximum likelihood and the QDA classifier takes on the form

$$a(\mathbf{x}) = \arg\max_{\mathbf{y}\in\mathbf{Y}}\left\{\ln(P_{\mathbf{y}}) - \frac{1}{2}(\mathbf{x} - \hat{\boldsymbol{\mu}}_{\mathbf{y}})^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\mathbf{y}}^{-1}(\mathbf{x} - \hat{\boldsymbol{\mu}}_{\mathbf{y}}) - \frac{1}{2}\ln[\det(\hat{\boldsymbol{\Sigma}}_{\mathbf{y}})]\right\}.$$

To provide a stable classification with the predictable error, it is necessary to control positive definiteness and well conditioning of $\boldsymbol{\Sigma}_{\mathbf{y}}$. It is important also to check possible significant disagreements with the multivariate normality of features for all classes. In this paper, we have used the Mardia test (see Mardia 1970) based on skewness and kurtosis measures.

Discriminant surfaces are all kinds of second-degree hypersurfaces in this case. In the cases of paraboloid and hyperboloid, we have extrapolation; that is, features located far from training samples correspond most likely to some other unknown group of objects. To use only closed surfaces restricting some finite area in the feature space, we supplemented the additional constraint

$$\max(p_{\mathbf{y}}) > P_{\min},$$

which allows us to introduce the special class of "unrecognized objects." In this case, all discriminant surfaces of the quadratic normal BC will delineate limited areas of feature space.

In practice, very often the multivariate Gaussian does not fit the true PDFs of features of one or several classes. In this case we can suppose that the family of distributions is not defined and construct empirical estimates of PDFs. A well-known nonparametric approximation of PDF is the kernel density estimation or Parzen–Rosenblatt window method that is a natural generalization of the normalized histogram. This estimate in the multidimensional case can be written as follows:

$$\hat{p}_h(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\prod_{j=1}^{M}\frac{1}{h_j}K\left(\frac{x^j - x_i^j}{h_j}\right),$$

where K is the kernel function (see Hastie et al. 2008) and h_j is the width of the window for the dimension j. In the general case, the choice of K does not essentially affect the classification accuracy. The estimated PDF has the same smoothness as the kernel function, thus smoother kernels should be used in case if corresponding smoothness of PDF is supposed.

On the contrary, the choice of h_j strongly affects the quality of the classification. Too small of a window width leads to overfitting PDF estimates and consequently to

TABLE 2. Parameters used for the estimation of classification quality.

Name of the characteristic	Abbreviation	Definition
Total classification error (overall error)	TE	$\mathrm{TE} = (1/N)\sum_{i=1}^{N}(y_i \neq \hat{y}_i)$
Total classification accuracy (overall accuracy)	TA	$\mathrm{TE} = 1 - \mathrm{TA} = (1/N)\sum_{i=1}^{N}(y_i = \hat{y}_i)$
Confusion matrix	CM	$\mathrm{CM}_{k,k} = \sum_{i=1}^{N}[(y_i = k)\,\&\,(\hat{y}_i = k)]$, $\mathrm{CM}_{k,j} = \sum_{i=1}^{N}[(y_i = k)(\hat{y}_i = j)], \quad (k \neq j)$
Omission error	OE	$\mathrm{OE}_k = \mathrm{OE}_k = \sum_{\substack{j=1 \\ (j \neq k)}}^{K} \mathrm{CM}_{k,j}$
Producer's accuracy	PA	$\mathrm{PA}_k = 1 - \mathrm{OE}_k$
Total omission error	TOE	$\mathrm{TOE} = (1/K)\sum_{k=1}^{K}\mathrm{OE}_k$
Commission error	CE	$\mathrm{CE}_j = \sum_{\substack{k=1 \\ (k \neq j)}}^{K} \mathrm{CM}_{k,j}$
User's accuracy	UA	$\mathrm{UA}_j = 1 - \mathrm{CE}_j$
Total commission error	TCE	$\mathrm{TCE} = (1/K)\sum_{j=1}^{K}\mathrm{CE}_j$
Kappa	κ	$\kappa = (P_o - P_e)/(1 - P_e)$, where $P_o = (1/K)\sum_{k=1}^{K}\mathrm{CM}_{k,k}$, $P_e = (1/K)\sum_{k=1}^{K}\mathrm{SC}_k^T \cdot \mathrm{SR}_k$, $\mathrm{SC}_j(1/K)\sum_{k=1}^{K}\mathrm{CM}_{k,j}$, and $\mathrm{SR}_k = (1/K)\sum_{j=1}^{K}\mathrm{CM}_{k,j}$

unstable classification working well only for training data. If the window is too large, PDF estimates become excessively smoothed and do not fit the true PDFs, leading to high classification errors. In this paper, the optimum window widths for each dimension of the feature space were found automatically from training data by using the leave-one-out cross-validation method.

5. Estimation of errors and regularization techniques

The performance of classification algorithms should be measured, adjusted and compared based on definite parameters calculated from the available dataset. The parameters used in this paper for the estimation of classification quality are represented in Table 2. The notation for the variables and parameters used is as follows: N is number of samples, K is number of classes, i is the index of the sample, j and k are the indices of classes, $\mathbf{y}$ are the reference (known) class names, and $\hat{\mathbf{y}}$ are the predicted class names. All of the performance parameters are calculated from elements of $\mathbf{y}$ (known classes) and corresponding elements of $\hat{\mathbf{y}}$ (predicted classes).

The total classification error (TE), which is also known as the overall error, is defined as the amount of incorrectly classified samples over the total number of samples. TE is the parameter that characterizes the quality of the classification on the whole without taking into account separate classes. The total classification accuracy (TA) is the amount of correctly classified samples over the total number of samples, and it can be calculated directly from TE.

The confusion matrix (CM) is the basic classification quality characteristics allowing a comprehensive visual analysis of different aspects of the used classification method. In this paper, we use the definition as follows: each row of CM represents a reference class; each column represents a predicted class. The vice versa definition is sometimes also introduced by some authors. CM is the basis for the calculation of a number of classwise quality parameters.

The omission error (OE) is defined as the probability of the false classification of features corresponding to some selected class. It can be calculated as the amount of false classified samples of the selected class over all samples of this class. The OE is referred to as a type-I error when taking the target class as a hypothesis. The producer's accuracy (PA) is the classwise accuracy from the point of view of the producer of reference classification (for instance, the expertise of in situ measurements). The PA is a complement of OE. The total omission error (TOE) is the mean OE over all classes considered. It is referred also as the classification error.

The commission error (CE) can be considered as the false alarm error for some selected class. It is defined as the probability of false classification for each possible classification result. The user's accuracy (UA) is the

TABLE 3. Number of measurements available for calibration and validation of algorithms and corresponding number of days.

	Total	Others	Breezes	Fogs	HWEs
Samples	42 431	40 556	1127	564	184
Days	442	441	43	20	10

accuracy from the point of view of the user, the person performing classification who would like to know the accuracy expected for each response of the classifier. The UA is a complement of CE. The total commission error (TCE) is the mean CE over all possible responses of the classifier used.

The kappa coefficient κ proposes to quantify the intensity of the actual agreement between the expert and automated classification results; κ can take values from -1 to 1. The zero value corresponds to completely random classification. It is considered that κ value can be classified as follows: $\kappa \in [1, 0.8)$ is excellent, [0.8, 0.6) is good, [0.6, 0.4) is moderate, [0.4, 0.2) is poor, [0.2, 0.0) is bad, and [0.0, -1] is very bad.

To calculate the presented above estimates, the block k-fold cross-validation technique was applied. Note that the standard random k-fold or hold-out cross validation with completely random sampling measurements is not enough objective test in the considered case. It should be noted that the random selection of training set leads to a large number of very similar samples of features in the testing set as the measurement frequency is 15 min. Thus, corresponding cross-validation estimates would be obtained in a dependent manner. For this reason, we used the block k-fold cross validation, which provides the independence of data in training and testing. The number of folds should not be too small because we have a risk in this case to exclude the whole season, which is the most important for the considered local events. Taking into account the number of measurements (see Table 3) and the duration of the campaign, we decided that the cross validation with 10 block folds is applicable in our case.

The cross-validation technique was used for estimation of classification quality is closely connected with one of the basic properties of learning classification algorithms—the generalization ability (GA). This concept means that prior, leave-one-out cross validation (Hastie et al. 2008) and independent estimates of the classification error do not have statistically significant differences. The GA of a learning algorithm indicates the important property of predictability of the accuracy, that is, when real classification errors are in agreement with their estimates.

The GA deals with the concepts of overtraining and curse of dimensionality. Overtraining is the undesirable phenomenon of the learning classification, when the discriminant surface depends on the high number of parameters that cannot be estimated well enough by training data. As the result, the discriminant surface well separates only the available training samples, but not the independent (test) samples. Thus, we have to use the most exact classifier providing the hypothesis of the equality of prior and cross-validation errors of classification.

The curse-of-dimensionality problem arises when by the increasing the dimensionality of the feature space we can

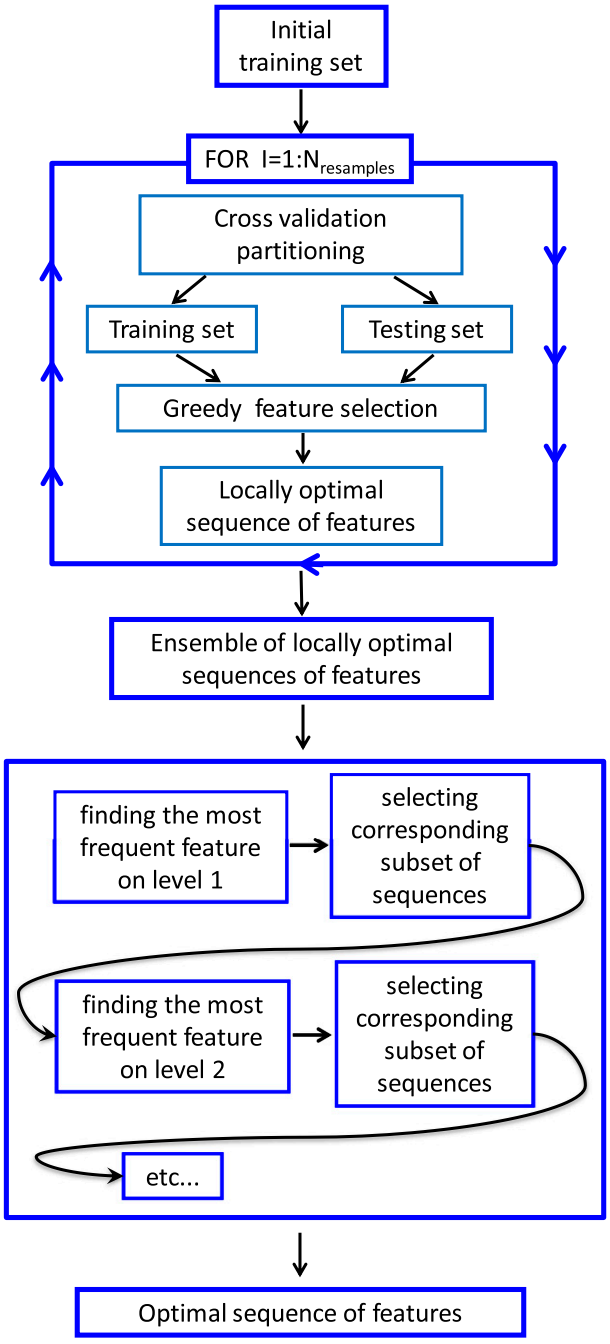

FIG. 2. Scheme of regularized greedy feature selection.

TABLE 4. Results of the regularized feature selection for different classification methods: optimized sequences of features and total omission errors for METEO experiments. Here, (*P*) signifies the possible set of features, when decreasing of TOE is finished and (*O*) signifies the feature corresponding to the last significant decreasing of TOE (optimal set).

	KNN, $N = 49$		SVM		QDA		PRW	
	Name	Error	Name	Error	Name	Error	Name	Error
1	HorWind	0.362	HorWind	0.359	HorWind	0.335	VWind	0.331
2	Hum	0.231	DirWind	0.252 (*O*)	Hum	0.22	UWind	0.249 (*O*) (*P*)
3	DirWind	0.164 (*O*)	Hum	0.224 (*P*)	VWind	0.165 (*O*)	DirWind	0.278
4	SolarRad	0.160	Rain	0.229	SigW	0.160 (*P*)	HorWind	0.310
5	SigW	0.159 (*P*)	SolarRad	0.259	Ustar	0.162	SigW	0.362
6	Rain	0.160	Ustar	0.314	MF	0.168	MF	0.413
7	T	0.163	MF	0.332	UWind	0.175	Ustar	0.456
8	HF	0.161	HF	0.349	HF	0.187	HF	0.505
9	Pres	0.169	SigW	0.355	SolarRad	0.202	Pres	0.641
10	MF	0.164	VWind	0.372	DirWind	0.214	Hum	0.703

achieve the exact classification of training samples; however, the classifier fails on the test set. In our case, the dimensionality of the predictor vector is up to 34. One of the solutions of this problem consists in the effective reduction of the feature space by selecting the most informative features. The greedy algorithm of stepwise forward selection is the standard and frequently used method of reduction of the feature space. It can be formulated as follows.

The features are divided into two groups—accepted in the classification model and remaining, for which an estimate of the possibility of acceptance into the model is checked. Features from the set of ''remains'' are consecutively added to the model and corresponding estimations of the classification error are calculated. We consider the TOE as the classification error, and not the TCE, because it is assumed that an expert will further process the data obtained from automated classification and it is more important not to omit the analyzed local events than to misclassify an ''others'' event (described below).

From the received set of errors, the minimum is chosen and compared with the error of the previous model. If a significant reduction of the error occurred, then the corresponding feature is accepted into the model, if not then the process stops. The disadvantage of this method consists in the instability of the obtained sequence of features. In practice, the selected features can significantly differ with small changes in the training set.

In this paper, we used the regularized greedy feature selection (Fig. 2). This method allows finding the more stable optimal sequence of features. Random partitioning of the learning dataset using the holdout cross validation is repeated many times. Then the stepwise forward selection is launched for each pair of the training and testing sets obtained. As a result, we have a set of locally optimal sequences of features. The sequences obtained as a rule are different in composition and length. The unique regularized solution of the feature selection problem can be obtained as the most probable sequence.

The algorithm of finding the most probable sequence of features consists in finding the mode on different levels (between firsts, seconds and the following members of sequences). In other words, we form a subset of the sequences corresponding to the mode (level 1) and

TABLE 5. Interval estimates of main classification quality parameters for METEO experiments.

Method	Quality parameter	Classification error with confidence intervals; possible set of features (*P*)	Classification error with confidence intervals; first two features
KNN	TOE	$0.152 < 0.159 < 0.182$	$0.217 < 0.234 < 0.261$
	TCE	$0.158 < 0.169 < 0.181$	$0.220 < 0.238 < 0.258$
	κ	$0.707 < 0.723 < 0.737$	$0.545 < 0.571 < 0.596$
SVM	TOE	$0.200 < 0.232 < 0.302$	$0.245 < 0.256 < 0.270$
	TCE	$0.162 < 0.176 < 0.196$	$0.217 < 0.228 < 0.241$
	κ	$0.666 < 0.702 < 0.724$	$0.601 < 0.617 < 0.631$
QDA	TOE	$0.153 < 0.161 < 0.169$	$0.213 < 0.221 < 0.229$
	TCE	$0.163 < 0.175 < 0.187$	$0.257 < 0.267 < 0.278$
	κ	$0.704 < 0.719 < 0.734$	$0.553 < 0.569 < 0.585$
PRW	TOE	$0.232 < 0.249 < 0.295$	$0.232 < 0.249 < 0.295$
	TCE	$0.258 < 0.278 < 0.304$	$0.258 < 0.278 < 0.304$
	κ	$0.551 < 0.581 < 0.597$	$0.551 < 0.581 < 0.597$

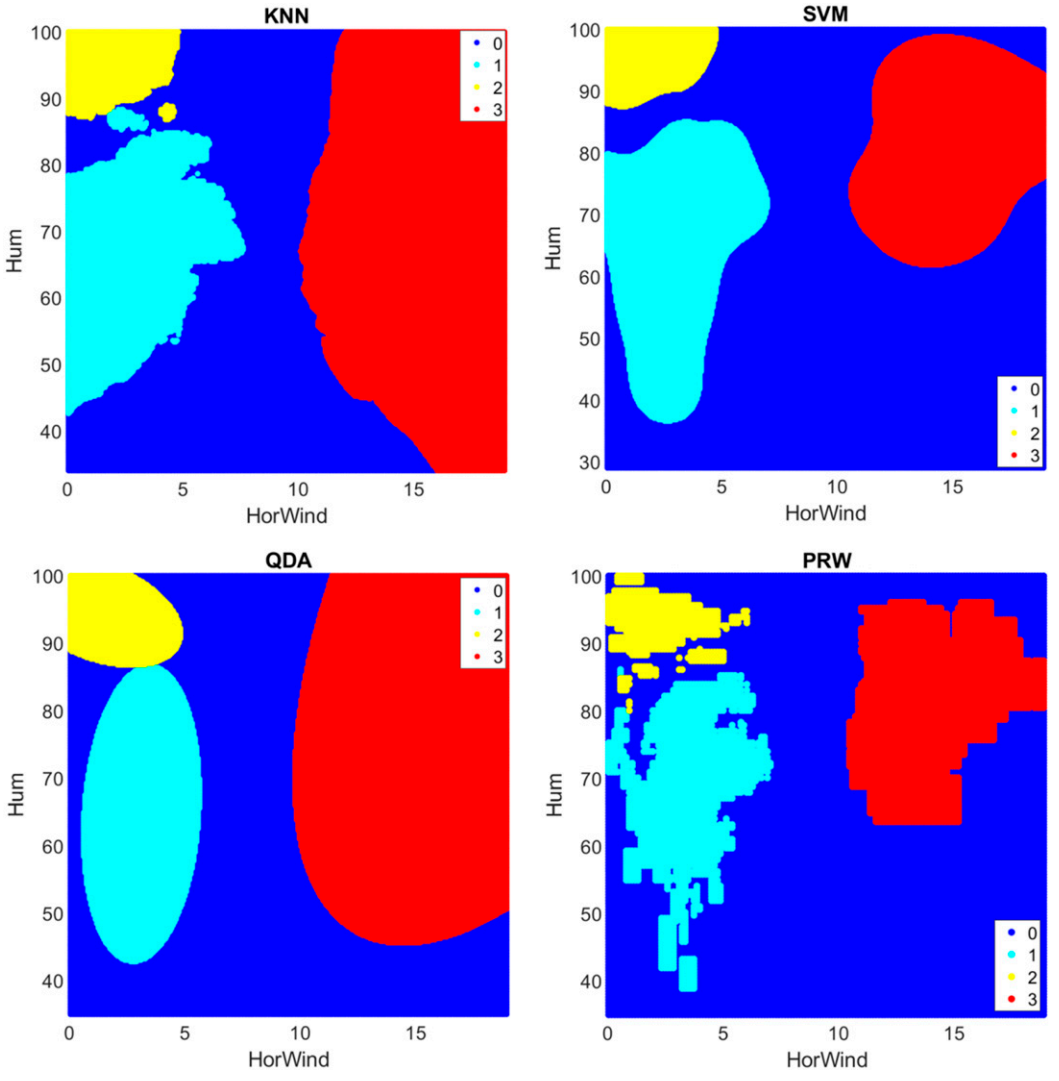

FIG. 3. Discriminant surfaces for the considered classifiers in HorWind–Hum feature space; 1000 samples of the others class are used. Here, 0 is others, 1 is breezes, 2 is fogs, and 3 is HWEs.

repeat the process with the next members. The process continues (levels 2, 3, and 4) up to selecting a unique sequence. In the case of a few equivalent sequences, the random choice is used. As a result, we can obtain the sequence of most informative features, which are robust to small changes in the training set.

Four classes were introduced for the dataset and expert classification described in sections 2 and 3: sea breezes, fogs, HWEs and other meteorological phenomena (hereinafter this class is called others). The classes are unbalanced as the number of measurements corresponding to the events classified is significantly different for the available dataset (see Table 3). This fact must be taken into account when training KNN and SVM classifiers, because it may affect the likelihood of recognizing the corresponding classes. In particular, as we can see, the total amount of breeze/fog/HWE events is relatively small, about 5%, in comparison with the others class. In the case of use of all of the available data for training, KNN and SVM classifiers will fail to correctly recognize the most part of local events, preferring to call them the prevailing others class.

To overcome this problem we used a reduced set of features corresponding to the others class with the number of samples approximately equal to the number of samples for breeze/fog/HWE events. Thus, for each classification method, we obtained a set of sequences of optimal features using random independent resampling of the others class and define the most probable optimal sequence using the method described in the previous section (see Fig. 2). This kind of optimization is also indispensable for the computationally expensive PRW method, as it diminishes the

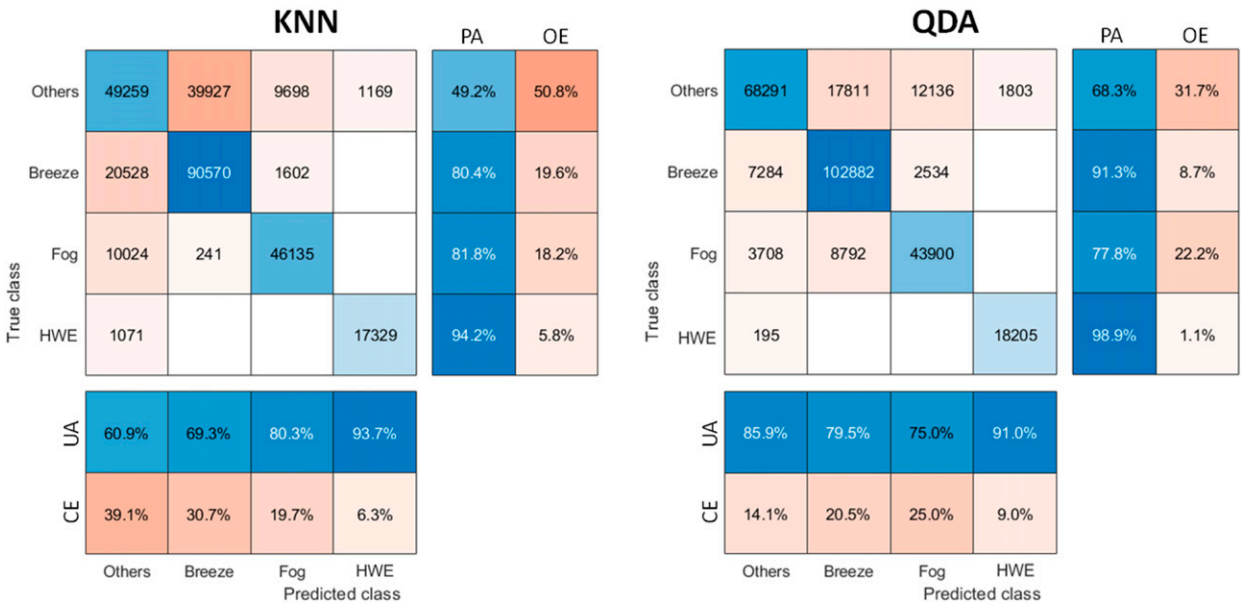

FIG. 4. Confusion matrices for KNN and QDA classifications using the HorWind and Hum predictors. Resampling was applied 100 times; 1000 samples of the others class are used. Corresponding omission errors and producer's accuracies are on the right side; commission errors and user's accuracies are on the bottom.

time of calculation dramatically. In numerical experiments for KNN, SVM, and QDA we used 49 different resampling of the others class. For PRW the choice of 11 resampling iterations is a reasonable compromise between the stability and the computational time.

To compare the performance of supervised ML algorithms we estimated confidence intervals for TOE, TCE, and κ, in addition to the point estimates of these parameters. Let α_e and β_e be point estimates of TOE, TCE, or κ, calculated for any two classifiers to be compared. Let $\alpha_l < \alpha_e < \alpha_u$ and $\beta_l < \beta_e < \beta_u$ be corresponding confidence intervals. The difference between α_e and β_e is not significant if $\alpha_e \in (\beta_l; \beta_u)$ or $\beta_e \in (\alpha_l; \alpha_u)$; otherwise the difference is significant. The confidence intervals and the elements of confusion matrices are estimated using a random resampling of the others class with 100 samples.

6. Results and discussion

Four algorithms of supervised classification described above were applied to the local meteorological data (the ultrasonic anemometer and the weather station) to recognize four types of local atmospheric events. The most probable optimal sequences were defined and classification errors were estimated.

a. Classification using meteorological dataset

In this set of numerical experiences (METEO), we used anemometer and meteorological station data for the classification. Parameters of the available dataset are presented in Table 3. Sequences of the most informative features (notations are presented in Table 1) selected for different ML algorithms are shown in Table 4. This sequence and corresponding classification errors are obtained using at the third step of the algorithm presented in the end of the introduction section.

Classification errors have similar behavior depending on the number of features used for all considered classifiers: the error decreases first, then stabilizes and further increases with some fluctuations. Lengths of decreasing sequences are different depending on classification method. However, the composition of first most informative components is very similar. For each classification method, we have fixed the components on which the error keep decreasing. This sequence is called the ''possible'' set of features, and the last element of the sequences is marked with (P) in Table 4. The possible set characterizes the most important parameters to archive the minimum of classification error on our dataset.

For such sets of features, we performed additional numerical experiments corresponding to the fourth step of algorithm indicated in the introduction section. Additional classification quality parameters TOE, TCE, κ, and corresponding confidence intervals were estimated, as described in section 5 and in Table 2. The results obtained for the possible set and for couples of the most informative predictors are shown in Table 5.

Another sequence of features, corresponding to the last significant decreasing of TOE, is called the ''optimal'' set, with the last element marked with (O). The

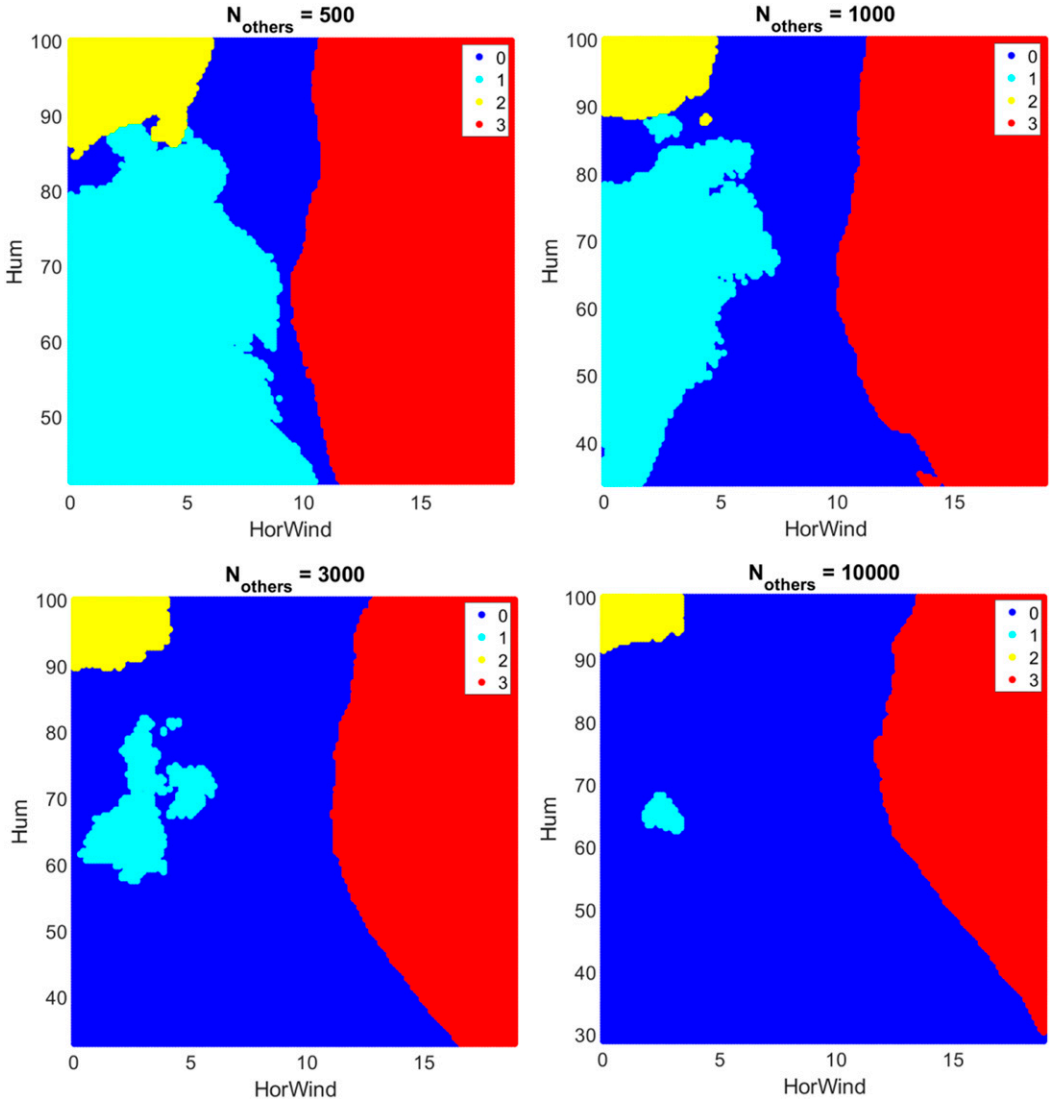

FIG. 5. KNN classification with different number of samples of the others class. Discriminant surfaces are constructed for 500, 1000, 3000, and 10 000 samples of others. Here, 0 is others, 1 is breezes, 2 is fogs, and 3 is HWEs.

optimal set is a subset of the possible set; we select the features until the decreasing TOE estimate enters into the confidence interval for TOE of the possible set presented in third column of Table 5. The optimal set of features takes into consideration only the parameters giving a significant decrease in TOE.

Possible (P) and optimal (O) sets of features were obtained in a similar way for all considered classification methods. The longest sequence of features (possible set) corresponds to KNN and consists of the five features (see Table 4). The TOE (classification error) is minimal for this method. Errors corresponding to first–second–third, first–second–third–fourth, and first–second–third–fourth–fifth features (calculated up to the DirWind, SolarRad, and SigW predictors in Table 4) are similar, and the difference between them is not significant, because they are inside the confidence interval of SigW TOE (0.152, 0.182) for a possible set (see TOE for KNN in Table 5, column 3). It means that the optimal set consist of three predictors: HorWind, Hum, and DirWind; we stop at third feature, because the DirWind TOE (0.164) is inside the confidence interval for a possible set (0.152, 0.182).

The interval estimates of TCE and κ for the optimal sequence are $0.158 < 0.169 < 0.181$ and $0.707 < 0.723 < 0.737$. As we can see from comparison of the third and the fourth columns of the Table 5, decreasing TCE and increasing κ are also significant for KNN between the two-predictor set and the possible set. The difference between TOE and TCE is not significant for the possible

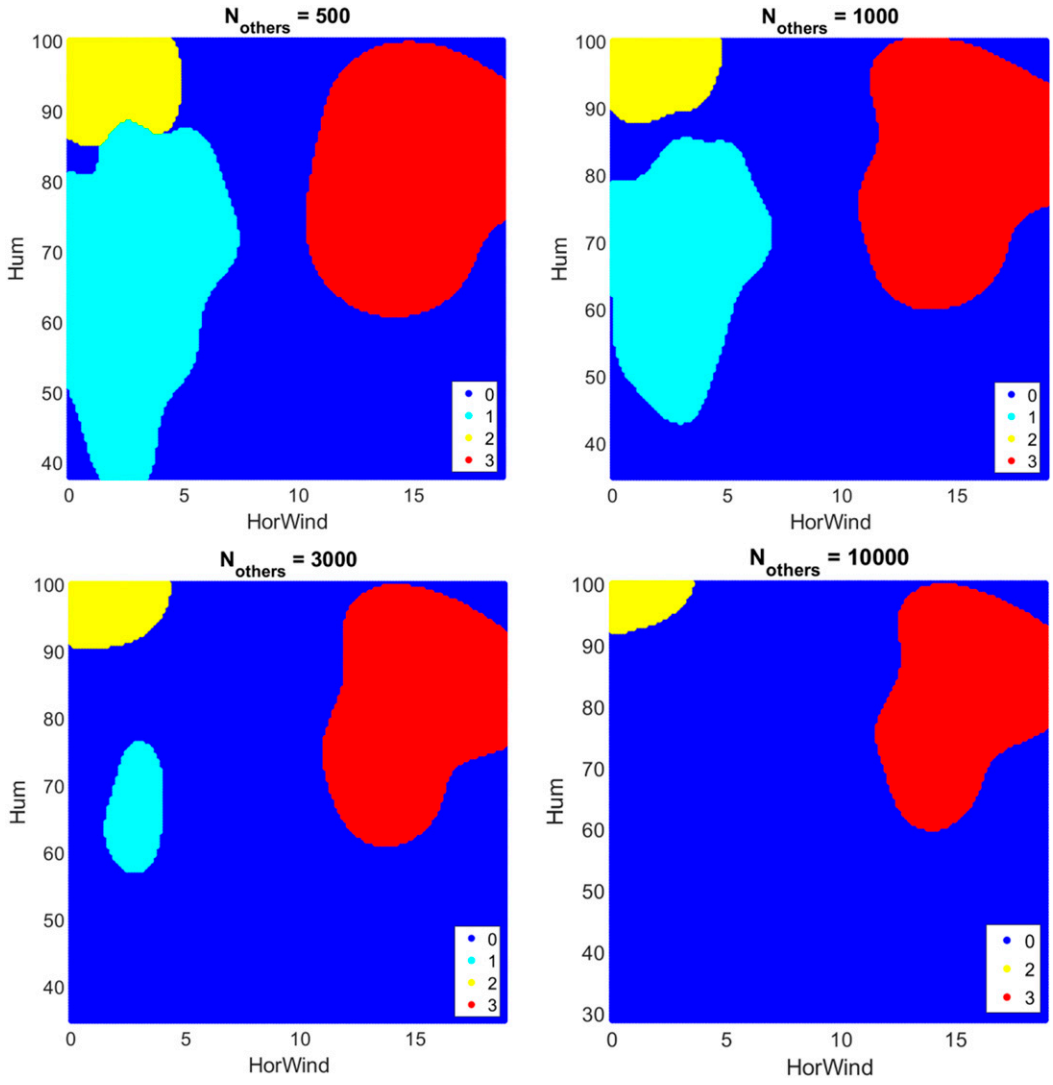

FIG. 6. As in Fig. 5, but for the SVM classification.

sequence for this classification method, as the estimate of TOE, 0.159 is in the interval for TCE, (0.158, 0.181), and, vice versa, estimation of TCE, 0.169, is in the interval for TOE, (0.152, 1.182). The value of κ reflects the good quality of the agreement of reference classes and classification results.

The QDA errors are similar to KNN. We can see from Table 5 that the difference is not significant.

Accuracy of the more sophisticated SVM and PRW methods is lower than that for KNN and QDA for both possible and optimal sequences. This is primarily due to the curse-of-dimensionality problem. As we can see from Table 4, TOE stops decreasing at the third element of sequences for SVM and at the second element for PRW. We can see from the Table 5, that the optimal sequence of PRW also contains two features. For the SVM method the difference between TOE and TCE is significant, as intervals for TOE and TCE in the third column of Table 5 do not intersect. Then, as we can see from the comparison of third and fourth columns of Table 5 for SVM, adding just one feature in the optimal sequence of SVM leads to a very strong expansion of the confidence interval of TOE from (0.245; 0.270) to (0.200; 0.302), so that it completely covers the previous interval. On the other hand, the difference between TCE for two-component (optimal) and possible sequences (fourth and third columns of Table 5) for SVM is significant; thus it is possible that these sequences coincide for SVM.

Discriminate surfaces for different classification methods are presented at Fig. 3 for HorWind and Hum. These surfaces divide two-dimensional predictor space to a few subspaces corresponding to classes. It

TABLE 6. As in Table 4, but for METEO+Avg2h experiments.

	KNN, $N = 49$		SVM		QDA		PRW	
	Name	Error	Name	Error	Name	Error	Name	Error
1	HorWind2h	0.353	HorWind2h	0.347	HorWind	0.344	VWind	0.329
2	Hum2h	0.230	DirWind2h	0.253 (*O*)	Hum2h	0.219	UWind	0.249 (*O*) (*P*)
3	DirWind2h	0.159 (*O*)	Hum2h	0.235 (*P*)	VWind	0.163 (*O*)	VWind2h	0.271
4	Rain2h	0.156	Rain	0.240	SigW	0.159 (*P*)	DirWind	0.307
5	MF	0.154 (*P*)	HorWind	0.246	VWind2h	0.159	HorWind	0.347
6	Rain	0.154	Rain2h	0.268	Hum	0.161	DirWind2h	0.371
7	HorWind	0.167	DirWind	0.289	MF	0.166	UWind2h	0.402
8	Hum	0.156	SolarRad	0.318	Ustar	0.168	HorWind2h	0.446
9	SolarRad2h	0.155	SolarRad2h	0.326	Rain2h	0.170	SigW	0.479
10	Pres	0.144	VWind	0.356	RH	0.177	MF	0.542

allows visualizing the classification by four supervised ML algorithms. The light-blue subspace is breezes, yellow are fogs, red are HWEs and the dark blue corresponds to the others class. The discriminant surfaces are defined at the learning step. Then, the classification of a new event by its predictor ($\text{HorWind}_{\text{New}}$, Hum_{New}) could be done visually. It is enough to see to which subspace the point belongs ($\text{HorWind}_{\text{New}}$, Hum_{New}).

KNN and PRW methods give the most complicated discriminant surfaces (see Fig. 3). Optimal sequences presented in Table 4 have the same length for KNN and QDA; the first two most informative predictors are also coincided: HorWind and Hum. The difference starts from the third parameter; however, it should be noted that in conjunction with the HorWind feature, VWind allows calculating DirWind; thus this difference in the third parameter is not very important.

Confusion matrices are presented in Fig. 4 for KNN and QDA methods using two features: HorWind and Hum. Through this couple of predictors the HWE class could be perfectly separated from the breeze and fog classes by both algorithms. In turn, the breeze and fog classes can also be discriminated well from each other, and corresponding errors are much less for KNN. The most significant errors appear for classification of other and breeze. In general, QDA classifier reveals better accuracy for these two features.

To illustrate the importance of the selection of the reasonable number of others events we plotted the discriminate surfaces for KNN and SVM classifiers at Figs. 5 and 6, respectively. We can observe for both methods the variation of areas corresponding to each class connected with the number of others events. The probability of classes 1–3 diminishes and the probability of class 0 (others) increases when the number of others samples increases. Note that the breeze class disappears for the SVM algorithm when the number of others events exceeds 10 000 samples (see Fig. 6). We considered 1000 others events to be a reasonable compromise to have comparable probabilities of each class. Bayesian QDA and PRW classifiers are less sensitive to the number of others events.

b. *Classification using meteorological and averaged meteorological parameters*

The analysis of misclassified events shows that ML algorithms fail for outliers of predictor occurring during a long-lasting event. For example, during a HWE, a decrease in the wind during a small period leads to misclassification of the event. To overcome this problem we suggested using moving average values of measured meteorological parameters as supplementary predictors. In addition, a contrast between measured and averaged parameters could bring some information about fluctuations of atmospheric measurements. We added, then, 2-h moving averages to predictor vectors. We use here abbreviation METEO+Avg2h to denote the second set of numerical experiments that uses anemometer data (see Table 1) and averaged values of meteorological parameters. The following 2-h-averaged parameters were added to the vector of predictors:

- mean horizontal wind HorWind2h,
- mean wind direction DirWind2h,
- mean wind west–east component UWind2h,
- mean wind south–north component VWInd2h,
- mean pressure Pres2h,
- mean rainfall Rain2h,
- mean solar radiation SolarRad2h
- mean temperature Temp2h, and
- mean relative humidity Hum2h.

Results of METEO+Avg2h numerical experiments are presented in Table 6. The decrease of classification error is observed with respect to previous experiment METEO (without averaged measurements; see Table 4) for KNN and QDA methods. Nevertheless, the analysis of confidence intervals (see Table 7) does not allow confirming that this decrease of classification error is significant. SVM and PRW methods do not take advantage of the availability of averaged values. Nevertheless, we note

TABLE 7. Interval estimates of main classification quality parameters for possible and optimal set of features for METEO+Avg2h experiments.

Method	Quality parameter	Classification error with confidence intervals; possible set of features (P)	Classification error with confidence intervals; optimal set of features (O)
KNN	TOE	$0.144 < 0.163 < 0.208$	$0.149 < 0.158 < 0.169$
	TCE	$0.148 < 0.166 < 0.187$	$0.151 < 0.163 < 0.175$
	κ	$0.708 < 0.730 < 0.749$	$0.718 < 0.734 < 0.750$
SVM	TOE	$0.209 < 0.245 < 0.341$	$0.240 < 0.259 < 0.288$
	TCE	$0.157 < 0.175 < 0.203$	$0.211 < 0.224 < 0.240$
	κ	$0.645 < 0.691 < 0.721$	$0.592 < 0.615 < 0.636$
QDA	TOE	$0.151 < 0.158 < 0.167$	$0.156 < 0.163 < 0.172$
	TCE	$0.159 < 0.171 < 0.182$	$0.180 < 0.189 < 0.205$
	κ	$0.710 < 0.724 < 0.739$	$0.696 < 0.713 < 0.727$
PRW	TOE	$0.232 < 0.249 < 0.295$	$0.232 < 0.249 < 0.295$
	TCE	$0.258 < 0.278 < 0.304$	$0.258 < 0.278 < 0.304$
	κ	$0.551 < 0.581 < 0.597$	$0.551 < 0.581 < 0.597$

that a few 2-h-averaged features are selected as important predictors in KNN, SVM, and QDA algorithms.

Confusion matrices for KNN and QDA algorithms are calculated (see Fig. 7) with a "possible" set of predictors (see Table 6). The analysis of Table 7 and Fig. 7 shows that these two algorithms have comparable performances and ensure the correct classification of 70%–80% of events. The comparison with the confusion matrix represented in Fig. 4 shows that the use of additional features changes the structure of the confusion matrix and improves the classification results. In this case, classification of HWEs remains the best besides classification errors have moved from others to fog. The accuracy of classification of breezes and fogs is slightly reduced; however, this is fully offset by an increase of classification accuracy of others events. The statistical estimates in the Table 7 show that the TOE difference between KNN and QDA is not significant.

c. *Computational aspects*

As classifications could often be executed independently, the algorithms were optimized for parallel calculations. Estimates of the overall calculation time for each classification method on a modern 4-core workstation are presented in Table 8. All numerical experiments were computed using parallel MATLAB software code on typical PCs/notebook computers except for PRW, which were executed on the "CALCUCLO" computational cluster. We reduced the number of samples from 49

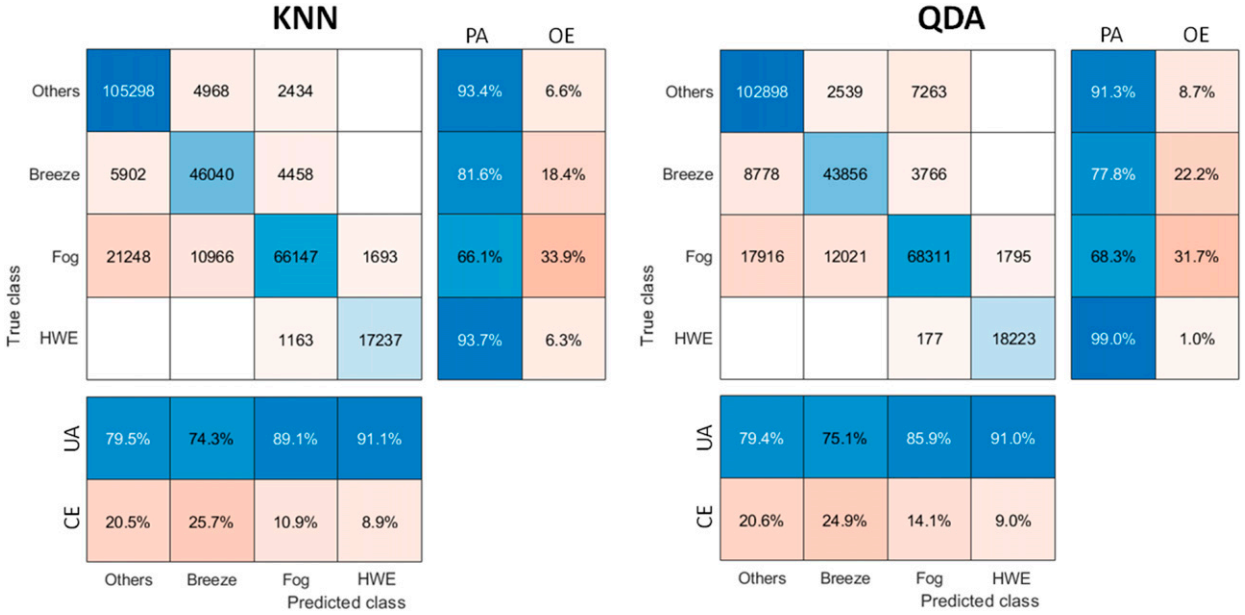

FIG. 7. As in Fig. 4, but for the METEO+Avg2h experiments.

TABLE 8. Estimated computation time (h) using typical 4-core Intel Core i7 workstation for four supervised machine-learning methods using 49 samples. Selection of features stops after choosing 10 locally optimal predictors.

Dataset	No. of predictors	No. of events	KNN, $N = 49$	SVM	QDA	PRW
METEO	13	42 431	0.2	8	0.1	500
METEO+Avg2h	22	42 431	0.3	10	0.2	700

to 11 to decrease the computation time for optimal feature selection of the PRW method. The most computationally complicated step is step 3 of the algorithm: the definition of the most informative predictor's sequence.

7. Conclusions and perspectives

Supervised machine learning algorithms allow one to correctly classify about 70%–80% of meteorological events. With the in situ anemometer data used to classify events, the most important predictors are the horizontal wind and humidity, then the wind direction, north–south wind component, and the solar radiation.

Among four supervised ML techniques, KNN and QDA algorithms have the best classification scores. SVM classifier has more important classification errors and PRW algorithm starts overlearning right after the selection of the second predictor.

The addition of averaged values of meteorological parameters to a feature vector allows the increasing of the KNN and QDA algorithms' precision, but the significance of this modification could not be shown yet.

To get better results, an expert should perform more mathematically accurate classification at the learning step. The latter classification corresponds to a larger scale and could categorize a meteorological event as HWE, even though it is a situation with a moderate wind.

Another important problem is that unlike a meteorologist, applied algorithms do not take into consideration the time dependence of meteorological parameters. For example, a specific evolution of a meteorological parameters combination (wind speed, humidity, temperature) is observed by a meteorologist for breeze detection. Thus, a time dimension of meteorological events should be taken into account by a more sophisticated technique than averaging.

An important limitation of the suggested algorithm is the absence of the visibility measurement for the fog detection. This kind of data could be available from Automated Weather Observing Systems or Automated Surface Observing Systems installed in local airports.

We expect that this method could also be applied to classify events by climatological in situ data or by outputs of meteorological or climate models. It would allow the estimation of meteorological events' frequencies in the perspective of climate change.

Acknowledgments. We gratefully acknowledge the financial support the Chemical and Physical Properties of the Atmosphere (CaPPA) project, which is funded by the French National Research Agency (ANR) through the Programme d'Investissement d'Avenir (PIA) under Contract ANR-11-LABX-0005-01 and by the Regional Council "Nord-Pas de Calais" and the European Funds for Regional Economic Development (FEDER). This research is a contribution to the CPER research project IRenE and CLIMIBIO. The authors thank the French Ministère de l'Enseignement Supérieur et de la Recherche, the Hauts-de-France region and the European Funds for Regional Economic Development for their financial support to this project. This work was also supported by the Russian Foundation for Basic Research (19-01-00215). Numerical experiments presented in this paper were carried out using the CALCULCO computing platform, supported by Service Commun du Système d'Information de l'Université du Littoral Côte d'Opale (SCoSI/ULCO). The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

Benkaciali, S., and K. Gairaa, 2014: Modélisation de l'irradiation solaire globale incidente sur un plan incliné. *Rev. Énerg. Renouvelables*, **17**, 245–252.

Boyouk, N., J.-F. Léon, H. Delbarre, P. Augustin, and M. Fourmentin, 2011: Impact of sea breeze on vertical structure of aerosol optical properties in Dunkerque, France. *Atmos. Res.*, **101**, 902–910, https://doi.org/10.1016/j.atmosres.2011.05.016.

Degefie, D. T., and Coauthors, 2015: Fog chemical composition and its feedback to fog water fluxes, water vapor fluxes, and microphysical evolution of two events near Paris. *Atmos. Res.*, **164–165**, 328–338, https://doi.org/10.1016/j.atmosres.2015.05.002.

Dietterich, T. G., and G. Bakiri, 1995: Solving multiclass learning problems via error-correcting output codes. *J. Artif. Intell. Res.*, **2**, 263–286, https://doi.org/10.1613/jair.105.

Ding, Y. H., and Y. J. Liu, 2014: Analysis of long-term variations of fog and haze in China in recent 50 years and their relations with atmospheric humidity. *Sci. China Earth Sci.*, **57**, 36–46, https://doi.org/10.1007/s11430-013-4792-1.

Dmitriev, E. V., V. V. Kozoderov, A. O. Dementyev, and A. N. Safonova, 2018: Combining classifiers in the problem of thematic processing of hyperspectral aerospace images. *Optoelectron. Instrum. Data Process.*, **54**, 213–221, https://doi.org/10.3103/S8756699018030019.

Doyle, M., and S. Dorling, 2002: Visibility trends in the UK 1950–1997. *Atmos. Environ.*, **36**, 3161–3172, https://doi.org/10.1016/S1352-2310(02)00248-0.

Dreyfus, G., 2005: *Neural Networks: Methodology and Applications*. Springer, 497 pp.

Dupont, J.-C., M. Haeffelin, A. Protat, D. Bouniol, N. Boyouk, and Y. Morille, 2012: Stratus-fog formation and dissipation: A 6-day case study. *Bound.-Layer Meteor.*, **143**, 207–225, https://doi.org/10.1007/s10546-012-9699-4.

Escalera, S., O. Pujol, and P. Radeva, 2010: On the decoding process in ternary error-correcting output codes. *IEEE Trans. Pattern Anal. Mach. Intell.*, **32**, 120–134, https://doi.org/10.1109/TPAMI.2008.266.

Friedman, J. H., J. L. Bentley, and R. A. Finkel, 1977: An algorithm for finding best matches in logarithmic expected time. *ACM Trans. Math. Software*, **3**, 209–226, https://doi.org/10.1145/355744.355745.

Gengembre, C., 2018: Multiscale variability of the coastal meteorology and aerosols under the influence of the industry. Ph.D. thesis, University of Littoral Cote d'Opale, 286 pp., http://www.theses.fr/2018DUNK0489.

Goswami, P., and S. Sarkar, 2015: Analysis and quantification of contrasts in observed meteorological fields for foggy and non-foggy days. *Meteor. Atmos. Phys.*, **127**, 605–623, https://doi.org/10.1007/s00703-015-0384-2.

Hastie, T., R. Tibshirani, and J. Friedman, 2008: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer Series in Statistics, Springer, 745 pp.

Iqbal, M., 1983: *An Introduction to Solar Radiation*. Academic Press, 408 pp.

Kambezidis, H. D., D. Weidauer, D. Melas, and M. Ulbricht, 1998: Air quality in the Athens basin during sea breeze and non-sea breeze days using laser-remote-sensing technique. *Atmos. Environ.*, **32**, 2173–2182, https://doi.org/10.1016/S1352-2310(97)00409-3.

Kessler, E., 1995: On the continuity and distribution of water substance in atmospheric circulations. *Atmos. Res.*, **38**, 109–145, https://doi.org/10.1016/0169-8095(94)00090-Z.

Liu, D. Y., S. J. Niu, J. Yang, L. J. Zhao, J. J. Lü, and C. S. Lu, 2012: Summary of a 4-year fog field study in northern Nanjing, Part 1: Fog boundary layer. *Pure Appl. Geophys.*, **169**, 809–819, https://doi.org/10.1007/s00024-011-0343-x.

Mardia, K. V., 1970: Measures of multivariate skewness and kurtosis with applications. *Biometrika*, **57**, 519–530, https://doi.org/10.1093/BIOMET/57.3.519.

Mavrakou, T., K. Philippopoulos, and D. Deligiorgi, 2012: The impact of sea breeze under different synoptic patterns on air pollution within Athens basin. *Sci. Total Environ.*, **433**, 31–43, https://doi.org/10.1016/j.scitotenv.2012.06.011.

Miller, S. T. K., Keim, B. D., Talbot, R. W., Mao, H., 2003. Sea breeze: Structure, forecasting, and impacts. *Rev. Geophys.*, **41**, 1011, https://doi.org/10.1029/2003RG000124.

Risi, C., V. Journé, J. Ricard, J.-Y. Grandpeix, and A. Spiga, 2016: Mise en évidence de la chaleur latente liée à l'évaporation et à la condensation de l'eau : Applications au fonctionnement des orages. *Meteorologie*, **94**, 15–18, https://doi.org/10.4267/2042/60700.

Saucier, W. J., 1955: *Principles of Meteorological Analysis*. University of Chicago Press, 438 pp.

Scholkopf, B., and A. J. Smola, 2002: *Learning with Kernels*. MIT Press, 626 pp.

Sen, Z., 2008: *Solar Energy Fundamentals and Modeling Techniques*. Springer, 276 pp.

Simpson, J. E., 1994: *Sea Breeze and Local Winds*. Cambridge University Press, 234 pp.

Sokolov, A., E. Dmitriev, E. Maksimovich, H. Delbarre, P. Augustin, C. Gengembre, M. Fourmentin, and N. Locoge, 2016: Cluster analysis of atmospheric dynamics and pollution transport in a coastal area. *Bound.-Layer Meteor.*, **161**, 237–264, https://doi.org/10.1007/s10546-016-0174-5.

Talbot, C., P. Augustin, C. Leroy, V. Willart, H. Delbarre, and G. Khomenko, 2007: Impact of a sea breeze on the boundary-layer dynamics and the atmospheric stratification in a coastal area of the North Sea. *Bound.-Layer Meteor.*, **125**, 133–154, https://doi.org/10.1007/s10546-007-9185-6.

Wang, Y., J. Zhang, A. R. Marcotte, M. Karl, C. Dye, and P. Herckes, 2015: Fog chemistry at three sites in Norway. *Atmos. Res.*, **151**, 72–81, https://doi.org/10.1016/j.atmosres.2014.04.016.

Xiang, Y., H. Delbarre, S. Sauvage, T. Léonardis, M. Fourmentin, P. Augustin, and N. Locoge, 2012: Development of a methodology examining the behaviours of VOCs source apportionment with micro-meteorology analysis in an urban and industrial area. *Environ. Pollut.*, **162**, 15–28, https://doi.org/10.1016/j.envpol.2011.10.012.

Ye, X., B. Wu, and H. Zhang, 2015: The turbulent structure and transport in fog layers observed over the Tianjin area. *Atmos. Res.*, **153**, 217–234, https://doi.org/10.1016/j.atmosres.2014.08.003.