Boundary-Layer Meteorology (2021) 179:133–161
https://doi.org/10.1007/s10546-020-00586-x

RESEARCH ARTICLE

Using Machine-Learning Methods to Improve Surface Wind Speed from the Outputs of a Numerical Weather Prediction Model

Naveen Goutham[1] · Bastien Alonzo[1] · Aurore Dupré[1] · Riwal Plougonven[1] · Rebeca Doctors[1] · Lishan Liao[1] · Mathilde Mougeot[2] · Aurélie Fischer[3] · Philippe Drobinski[1]

Received: 8 March 2019 / Accepted: 29 October 2020 / Published online: 14 January 2021

Abstract
The relationship between the wind speed derived from the outputs of a numerical-weather-prediction model and from observations is explored using statistical and machine-learning models. Eight years of wind-speed measurements at a height of 10 m (from 2010 to 2017) from 171 stations spread over mainland France and Corsica are used for reference. Operational analyses from the European Center for Medium Range Weather Forecasts (ECMWF) provide the model information not only on the surface flow, but on other aspects of the atmospheric state at the location (or above) each station. In a first step, a large number of explanatory variables are used as input to several models (linear regressions, k-nearest neighbours, random forests, and gradient boosting). The modelled wind speed in the ECMWF analyses, by itself, has root-mean-square errors over all stations distributed widely around a median of 1.42 m s^{-1}. Using statistical post-processing and making use of a historical dataset for training, the median of the root-mean-square errors at all stations can be reduced down to 1.07 m s^{-1} when modelled with linear regressions, and down to 0.94 m s^{-1} with the machine-learning models (random forests or gradient boosting). Yet more significant decreases are found for coastal stations where the errors are largest. The random-forest models are further explored to reduce the list of explanatory variables: a list of 25 explanatory variables, mainly consisting of flow variables (wind speed, velocity components, horizontal gradients of geopotential on different isobaric surfaces, wind shear between 10 and 100 m) and including marginally some temperature variables, appears as a good compromise between performance and simplicity. Finally, as a preliminary test for further work, the relation thus captured between the model outputs and the observed wind speed at a given time is applied to forecasts of the numerical-weather-prediction model, for lead times up to 24 h. The machine-learning model is found

✉ Riwal Plougonven
riwal.plougonven@lmd.polytechnique.fr

1 Laboratoire de Météorologie Dynamique/IPSL, Ecole Polytechnique, CNRS, Palaiseau, France

2 Department of Applied Mathematics, Ecole Nationale Supérieure d'Informatique pour l'Industrie et l'Entreprise, Evry, France

3 Laboratoire de Probabilités et Modèles Aléatoires, Université Paris Diderot Paris 7, Paris, France

to be essentially as relevant on the forecasts as it was on the analyses, encouraging further use and development of these approaches for local wind-speed forecasts.

Keywords Downscaling · Machine learning · Surface wind speed

1 Introduction

Surface wind speed is a meteorological variable of considerable importance because it affects human activities in a number of ways, including damage to buildings, fallen tower cranes, and injuries due to objects carried by strong winds. Over the past decade, the significant development of wind energy has created a new motivation and demand for estimations of the wind speed near the surface. Notably, the evolution of regulations for the pricing of wind energy (from feed-in tariffs to market prices) implies an increased demand for accurate forecasts of surface wind speed at wind-farm locations.

Numerical-weather-prediction (NWP) models constitute a major source of information on surface flows. However, as surface flows are turbulent and strongly influenced by small-scale features absent in the limited representation of NWP models, the modelled surface velocity components, when compared with local observations at a given site, generally contain large errors, including biases. For a given site where observations are available for a long enough interval, it is logical to use these observations to learn from and correct the biases and errors of the model for that location. In fact, estimating a local quantity based on output of a NWP model and past observations at a given location has been an active field of research for half a century, called model output statistics (Glahn and Lowry 1972). Glahn and Lowry (1972) have applied multilinear regressions to several variables, including surface wind speed, using a forward-stepwise-screening procedure to select the variables used as predictors. Nowadays, it is common for operational centres to use model output statistics to provide forecasts of quantities where observations are available (Wilson and Vallée 2002; Baars and Mass 2005; Schmeits et al. 2005; Kang et al. 2011; Zamo et al. 2014). As the methodology of weather forecasts evolved, from deterministic to probabilistic, some of the approaches used for model output statistics have also changed (Schuhen et al. 2012).

Fundamentally, the endeavour to estimate a small-scale, unresolved, fluctuating quantity from modelled knowledge of the large-scale field connects with several research fields with different aims, different sources of information, and different criteria for validation. One is model output statistics, which generally focuses on a given location for which observations are available. Another related activity is downscaling, i.e. elaborating a procedure to estimate a variable sensitive to small scales based on information on the large-scale flow. When used in the context of climate projections, the aim is to generate plausible time series of local variables in climate-change scenarios as proposed, for example, with the Statistical Down-Scaling Model (Wilby and Dawson 2013). Downscaling has been applied to estimate surface velocity components with an emphasis on identifying variables which carry information (Salameh et al. 2009; Devis et al. 2013). For locations in southern France, where topographic effects crucially affect the flow, Salameh et al. (2009) used generalized additive models to simulate small-scale velocity components from the reanalyses of the European Centre for Medium-Range Weather Forecasts (ECMWF).

Finally, the need to estimate subgrid-scale components of the velocity from modelled knowledge of the large-scale flow motivates the development of parametrizations in weather and climate models (e.g., Kalnay 2003). These differ in profound ways, seeking a generic

relation between the large-scale flow and the effects of unresolved small-scale components of the flow. There is, to our knowledge, little exchange between research on parametrizations and research on downscaling. Nonetheless, there may be opportunities to learn: for instance, downscaling studies provide bounds on the fraction of the local, subgrid-scale signal that can be reconstructed from knowledge of the large-scale flow, and on the relative importance of explanatory variables that contribute to this reconstruction.

The present study has common aims with model output statistics or downscaling, i.e. improving the determination of low-level wind speed, at locations where observations are available, using information from a NWP model and statistical/machine-learning models trained on past observations. For a given location where historical wind-speed measurements are available, the comparison of the measurements to NWP outputs is bound to show some significant errors, some of which one may hope to reduce while others should be expected to remain (de Rooy and Kok 2004). The sources of errors can be identified as:

- model error: the model describes the atmospheric flow only approximately, partly because of limited resolution, partly because processes that occur at small scales are represented through parametrizations.
- representativity error: the modelled value represents some average over space. For a variable such as 10-m wind speed having many small-scale variations (those due to turbulence may average out in time, but those due to local effects, such as roughness inhomogeneity and obstacles, do not necessarily), a local value is bound to differ from the value for a grid box (e.g., Horlacher et al. 2012).
- predictability limits (when considering forecasts): even if the model is perfect, errors, however small in the initial states, grow in forecasts because of the chaotic nature of the atmospheric flow. For short lead times of a day or less, this should be a minor source of error (Kalnay 2003).

The skill of NWP models is continuously increasing (Bauer et al. 2015), as are their spatial resolutions. Both of these facts imply that the description of 10-m wind speed by models is improving. Surface velocity components that are directly output from numerical weather prediction models still include significant errors (Haiden et al. 2018), while other variables such as pressure, which varies on considerably larger scales, are more accurate.

The question of precisely estimating the wind speed at specific locations has received renewed interest from the wind-energy sector. Very different approaches have been considered for forecasting the wind speed at locations of wind farms for different lead times: for short lead times of minutes to a few hours, statistical/machine-learning models trained with the locally observed wind speed have been developed using a variety of techniques (e.g., Chang 2014; Tascikaraoglu and Uzunoglu 2014; Foley et al. 2012; Wang et al. 2011). For longer lead times, from half a day to several days, outputs from NWP models have been used, including model output statistics for wind speed (Ranaboldo et al. 2013; Lazic et al. 2014) and of solar irradiance (Mejia et al. 2018). The most common practice in these cases remains the use of linear or multilinear regression, with a central issue being the choice of explanatory variables. Ranaboldo et al. (2013) present a stepwise-screening procedure to identify the most relevant variables to forecast 10-m wind speed at two locations, showing that variables describing the flow lead to the best performances.

The purpose here is to explore and improve the estimation for the local 10-m wind speed from recent outputs of the ECMWF model over stations in France sampling different geographical settings. Specific issues considered are the performance of the NWP model and the improvement gained by using parametric and non-parametric models. More precisely, the emphasis is put on evaluating the improvement gained by the use of machine-learning

models. Another objective is to identify those variables in the NWP model output that carry the most information to reconstruct the surface wind speed.

This study builds on the exploration of parametric and non-parametric models of surface wind speed introduced in Alonzo et al. (2018). Measurements of the velocity components were used for one specific location, and a detailed comparison of regression models was carried out for that particular site. The best performance was obtained with linear regression, after an appropriate selection of input variables. Random-forest models performed nearly as well without the need for a detailed expertise. Here, we extend this first work to more than 150 stations over France, making it possible to test the performance of different parametric and non-parametric models in several geographical contexts.

The data and methods used are described in Sect. 2. The performance of the numerical weather prediction model and of its combinations with different post-processing models are assessed and compared in Sect. 3. Focusing on the best model, we reduce the number of explanatory variables and identify what seems, for all stations, to constitute the most informative list of variables (Sect. 4). Other aspects and issues, such as the diurnal cycle, are discussed in Sect. 5. Before concluding, we show for one station that the improvements gained from training with past observations and analyses also carry over to forecasts (Sect. 6).

2 Data and Methodology

The data used includes wind-speed observations and outputs from a NWP model, as described below (Sect. 2.1). The statistical models used are described in Sect. 2.2, and the procedure to separate the data into training and testing datasets is explained at the end of that section.

2.1 Data

The Integrated Surface Database is a global database of meteorological observations available at an hourly time resolution (Smith et al. 2011). About 400 weather stations in France provide their observations in the Integrated Surface Database. In order to better train the models, we retained stations with over 90% of available data for a span of 8 years from 2010 to 2017. The retrieved observed data comes from 171 stations well distributed across mainland France and Corsica.

The ECMWF is an intergovernmental operational centre that provides medium-range weather forecasts on a global scale, and has the largest repository of archived global weather data. Operational analyses[1] from the ECMWF are retrieved with a spatial resolution of 0.125° in latitude and longitude over mainland France and Corsica. While this is a fine resolution for a global NWP model, this remains coarse-grained when comparing 10-m wind speed to measurements at one specific location, given for instance the sensitivity to the local topography.

The local surface wind speed is related to the synoptic-scale flow. The large-scale (synoptic) systems like depressions, fronts, and storms are described in terms of physical variables at different pressure levels including wind speed, geopotential height, divergence, vorticity, and temperature (Table 1). However, the intra-day wind-speed variations that occur in the boundary layer may not be wholly explained by the synoptic flows. The variables that convey information about the stability of the boundary layer include, but are not limited to, the

[1] Estimate of the atmospheric state at any given time obtained by assimilating observed data from within a time window around the corresponding time to previous forecasts made by the NWP model

Table 1 Explanatory variables from the interior of the NWP model domain, retrieved on pressure levels

Pressure level (hPa)	Variable	Unit	Symbol
1000/925/850/500	Zonal velocity component	$m\ s^{-1}$	u
1000/925/850/500	Meridional velocity component	$m\ s^{-1}$	v
1000/925/850/500	Geopotential height	$m^2\ s^{-2}$	z
1000/925/850/500	Divergence	s^{-1}	d
1000/925/850/500	Vorticity	s^{-1}	vo
1000/925/850/500	Temperature	K	T

Table 2 Explanatory variables retrieved from surface variables of the NWP model. The last three variables are accumulated over the last 6 hours

Altitude	Variable	Unit	Symbol
10 m/100 m	Wind speed	$m\ s^{-1}$	F
10 m/100 m	Zonal velocity component	$m\ s^{-1}$	u
10 m/100 m	Meridional velocity component	$m\ s^{-1}$	v
2 m	Temperature	K	$t2m$
surface	Skin temperature	K	skt
–	Mean sea level pressure	Pa	msl
surface	Surface pressure	Pa	sp
–	Boundary-layer height	m	blh
–	Boundary-layer dissipation	$W\ m^{-2}$	bld
surface	Surface latent heat flux	$W\ m^{-2}$	$slhf$
surface	Surface sensible heat flux	$W\ m^{-2}$	$sshf$

Table 3 Explanatory variables computed as differences in the vertical between two height or pressure levels

Vertical level	Variable	Unit	Symbol
10 to 100 m	Bulk wind difference	$m\ s^{-1}$	DF
1000 to 925 hPa	Bulk wind difference	$m\ s^{-1}$	DFP
1000 to 925 hPa	Temperature difference	K	DTP

temperature, heat fluxes, the surface pressure, and the boundary-layer dissipation (Table 2). These variables at the grid points are referred to as raw data hereafter. Other important variables that convey information about the vertical-exchange processes in the boundary layer are the vertical wind shear and the temperature gradient which are computed from the raw data as shown in Table 3.

The main set of quantities to be used in the parametric and non-parametric models for a specific station is obtained from the bi-linear interpolation of data at the closest grid points of the ECMWF model surrounding that station. We also computed a set of four quantities by taking north–south, east–west, and diagonal gradients around each station, estimated using finite differences. We found that the north–south and east–west gradients were more significant than the diagonal gradients. Hence, for each quantity, we retained its value interpolated at the station location and the two components of its gradient (north–south and east–west) as

explanatory variables to feed into the machine-learning models. This leads to 117 explanatory variables for each station.

The time period covered by the dataset is April 2010–December 2017. In order for the observed data to match the 6-h frequency of the ECMWF model outputs, we defined a 2-h averaging window by only considering the observed data at the hour, an hour before and after the top of the hour, at 0000, 0600, 1200 and 1800 UTC.

The ability of the ECMWF model to represent the observed wind speed is quantified by the root-mean-square error (r.m.s.e.) denoted by $E_{w,obs}$, and Pearson's correlation coefficient $\rho_{w,obs}$, given in Eqs. 1 and 2 respectively, where w stands for the time series from the ECMWF analyses, and obs for the observed wind speed

$$E_{w,obs} = \sqrt{\frac{\sum_{t\in\mathcal{S}} \left(y_t^w - y_t^{obs}\right)^2}{|\mathcal{S}|}}, \tag{1}$$

$$\rho_{w,obs} = \frac{\sum_{t\in\mathcal{S}} \left(y_t^w - \bar{y}^w\right)\left(y_t^{obs} - \bar{y}^{obs}\right)}{\sqrt{\sum_{t\in\mathcal{S}} \left(y_t^w - \bar{y}^w\right)^2}\sqrt{\sum_{t\in\mathcal{S}} \left(y_t^{obs} - \bar{y}^{obs}\right)^2}}, \tag{2}$$

where $\mathcal{S}$ denotes the set of indices of the data, with $|A|$ the number of elements of a set A, and $\bar{y} = \frac{1}{|\mathcal{S}|}\sum_{t\in\mathcal{S}} y_t$ the mean of the time series y.

Figure 2 shows the r.m.s. error and correlation coefficient for the 10-m wind speed between the observations and the ECMWF analyses, for all the meteorological stations under consideration in France. Figure 2a shows that the r.m.s. error of the wind speed from the ECMWF analyses exceeds 1 m s^{-1} for most of the inland stations: the minimum at an individual station is 1 m s^{-1}, the maximum is 4.6 m s^{-1}. The average over all stations is 1.7 m s^{-1}, with a standard deviation of 0.8 m s^{-1}. The coastal stations in the west, south and Corsica have a higher r.m.s. error greater than 2 m s^{-1}. In Fig. 2b, we see that the correlation coefficient for inland stations in the north is about 0.8, whereas for stations in the south and along the coasts it hardly reaches 0.7 and can be as low as 0.4. Note that, because of the higher r.m.s. error and the lower correlation coefficient found along the coasts, special attention was paid to these stations during interpolation to check if the location of grid points from the model has an effect. Upon careful examination, it was noticed that the location of grid points has no significant influence. The poorer performance may be due to factors that likely contribute to the difficulty of modelling wind speed at the coast. These include the discontinuity in surface conditions and the ensuing complexity of the boundary layer, and also possibly local phenomena such as the sea breeze.

We computed the annual averages of the r.m.s. error and correlation coefficient of the ECMWF analyses over all stations (see Fig. 1). An improvement of the performance of the model in the year 2014 is observed (a decrease of the r.m.s. error) resulting from changes in the ECMWF model, notably a modification of the parametrization of surface drag and the increase of the vertical resolution from 91 to 137 levels in June 2013 (Riddaway 2013). Nevertheless, these changes did not have an impact on the correlation coefficient. The average r.m.s. error and the correlation coefficient for the time period 2010–2017 are 1.7 m s^{-1} and 0.68 respectively. It is also instructive to include the median of the r.m.s. error for all stations: 1.4 m s^{-1}. This value is smaller than the average, as can be expected for a positive variable which can be very large in locations where the model performs very poorly. These errors are significant given that the time-averaged wind speed averaged over all stations is 3.4 m s^{-1}. More precisely, we calculated the ratio of the r.m.s. error to the time-averaged wind speed for each station. The overall mean of these ratios is 0.52, implying that any significant decrease of the error is worthwhile.

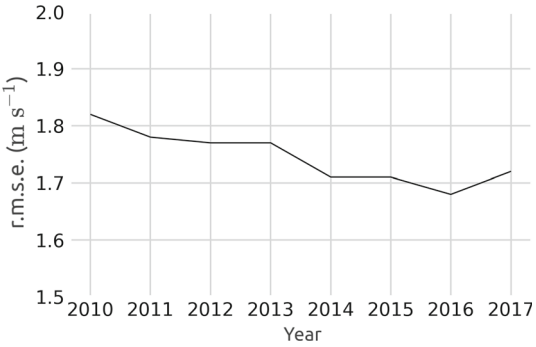

Fig. 1 Root-mean-square error of the raw output of the ECMWF analyses for the wind speed at 10 m over all the stations in France for the years 2010 to 2017. Extreme values are 1.82 (in 2010) and 1.68 m s^{-1} (in 2016), and the average is 1.74 m s^{-1}. The correlation coefficient is stable during this period and equal to 0.68 except for 2 years (it is 0.67 in 2010 and 2013)

2.2 Methodology

The aim here is to model the 10-m wind speed at the above-mentioned meteorological stations in France from the outputs of the ECMWF model, building on the work of Alonzo et al. (2018). Here, the target variable is the observed wind speed and explanatory variables derive only from output of the ECMWF model ($p = 117$ explanatory variables).

In statistics and machine learning, two main classes of methods may be used: parametric or non-parametric methods. In a parametric model, the relationship between outputs and inputs may be described analytically based on some probability distribution (for instance, a Gaussian model). On the contrary, a non-parametric method does not rely on a particular distribution assumption for the data, but involves several tuning parameters.

2.2.1 Linear Regression

Linear regression is a widely used model, which identifies a linear relationship between the response Y_t and the explanatory variables $X_t^1, X_t^2, \ldots X_t^p$ at a given time t

$$Y_t = \beta_0 + \sum_{j=1}^{p} \beta_j X_t^j + \varepsilon_t, \tag{3}$$

where the coefficients β_j are the regression coefficients estimated using a least-squares approach, and ε_t is the error.

For a large number of variables, in order to obtain a precise estimation, it is necessary to select the most relevant variables. Many methods are available, either forwards or backwards, to retain only a subset of the explanatory variables. Forwards selection starts with an empty list of predictors adding one highly significant predictor at each step until a stopping criterion is reached, whereas backwards selection starts with a full list of predictors eliminating one highly insignificant predictor at each step until a stopping criterion is reached. Omitting the Gaussian assumption, Lasso regression (also called ℓ^1 regularization) may be employed to select the most important predictors by adding a penalty term to the least-squares error (Gareth et al. 2013; Tibshirani 1996). This penalty acts as a constraint favouring a weaker sum of the absolute values of the regression coefficients; this results in some of the coefficients reducing to zero, implying that the corresponding explanatory variable is dropped.

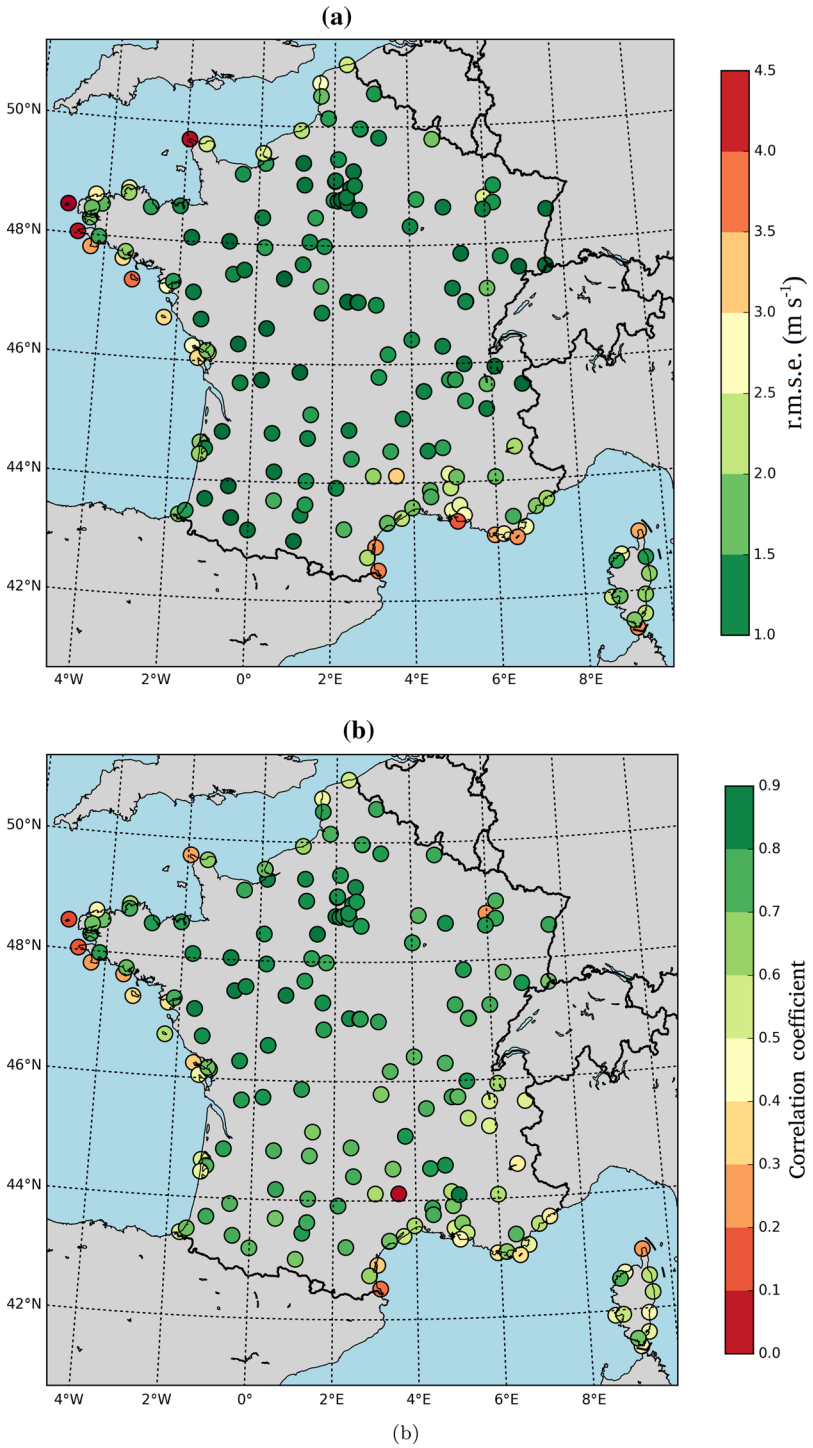

Fig. 2 Root-mean-square error of 10-m wind speed in the ECMWF analyses (**a**); correlation of 10-m wind speed in the ECMWF analyses (**b**)

2.2.2 Random Forests

In non-parametric frameworks, decision trees are today commonly used for modelling. Decision trees split iteratively the input space by minimizing the target variance on each side of the split (Muller and Massaron 2016). Decision trees have the advantage of being easy to set up and understand, and can capture nonlinear relations between the explanatory variables and the target. Training a single decision tree on a dataset would however lead to overfitting. Moreover, decision trees may suffer from a large variance: if the training dataset is split into two parts, and if a decision tree is fit for each of the two halves, the results could be quite different. To remedy this, bagging (bootstrap aggregation) consists of drawing multiple subsets for training the model (bootstrap), and then aggregating together the resulting trees. The resulting variance is correspondingly lower, the risk of overfitting is much reduced. This method has been demonstrated to significantly enhance accuracy, and can be further improved by an additional modification in the procedure leading to a random forests model (Breiman 2001). A random forest is an ensemble of many regression trees built with a random selection of the features used for each split, to decorrelate the different trees and further reduce the risk of over-fitting to the training dataset. The prediction is given by the average of all the leaf-response values in the training dataset. The random-forest parameters are the number of trees in the forest (100 for this application) and the proportion of explanatory variables allowed at each split (here, the square root of the number of variables). Finally, boosting grows trees sequentially by specially updating the weights of the worst predicted observations. It consists of using the information from the errors of previously obtained trees, and slowly learning to reduce those errors. This learning method, when used with gradient-descent optimization, is named gradient boosting. The boosting parameters are the number of trees (here, 100), and depth of the individual trees (here, 10).

Random forests were chosen as our main tool for exploring the potential of non-parametric models because they have been demonstrated to be efficient (Gareth et al. 2013), are based on decision trees, which are fairly easy to understand, and are interpretable: by counting how frequently one explanatory variable is used to define a split of the data into two subsets, it is possible to evaluate the relative importance of the different explanatory variables. In other words, random forests inform us about the information content of the different explanatory variables relative to our target. Whether the non-parametric method is retained for further use or not, this information in itself is of great value. Such information is not available from artificial neural networks.

2.2.3 Nearest Neighbours

An alternative to tree-based methods may be the k-nearest neighbour algorithm, which takes the k closest training observations based on the Euclidean distance and predicts the output as the average of the outputs of the k-nearest neighbours. Note that k is in this case a crucial parameter to tune. This model is retained as an alternative and cheap non-parametric model, and for its great simplicity.

2.2.4 Training and Validation

In order to train and test the different machine-learning models, 10-fold cross-validation is used. This procedure splits the data into a training dataset, and a dataset to test and evaluate the perfomance of the model (Gareth et al. 2013; Alpaydin 2010). Our dataset is partitioned into

Table 4 Parametric and non-parametric models implemented in this work

Machine-learning model	Name
Linear regression with all variables	LRall
Linear regression with stepwise selection	LRstep
OLS with Lasso regularization	Lasso
Random forest with all variables	Random
Gradient boosting with all variables	Gradient
k-nearest neighbour with the best 10 chosen variable from Random	k-Nearest

10 subsets, keeping a tenth of the dataset for verification every time, as is common practice with large datasets. Training is performed in a cyclic way on nine subsets keeping the last one to evaluate the model. Global performances are computed by averaging the 10 repetitions. The Python packages used are NumPy, SciPy, matplotlib, pandas, and Scikit-Learn.

3 Comparison of Different Parametric and Non-Parametric Models

The performance of the machine-learning models relative to ECMWF raw model output is computed using a relative error for the r.m.s. error and the correlation coefficient (given as a percentage)

$$\Delta E_{ML} = \frac{(E_{ECMWF} - E_{ML})}{E_{ECMWF}} 100\%, \tag{4}$$

$$\Delta \rho_{ML} = \frac{(\rho_{ML} - \rho_{ECMWF})}{\rho_{ECMWF}} 100\%, \tag{5}$$

where E_{ECMWF} and ρ_{ECMWF} are the r.m.s. error and correlation coefficient computed between the ECMWF model and the observations; E_{ML} and ρ_{ML} are the r.m.s.e. and correlation correlation computed between the predictions of a given machine-learning or statistical model and the observations.

The parametric models implemented are linear regression with all explanatory variables (hereafter LRall), linear regression with stepwise selection of variables (hereafter LRstep), and the ordinary least squares with Lasso regularization (hereafter Lasso). The non-parametric models implemented in this work are random forest with all variables (hereafter Random), gradient boosting (hereafter Gradient), and k-nearest neighbours (hereafter k-Nearest) using the 10 most important explanatory variables provided by the Random model.

All models are summarized in Table 4. Two more k-nearest-neighbours models were also trained but are not mentioned because of their poor performances: one with all explanatory variables and another with only five explanatory variables related to the velocity.

3.1 Performance of the Machine-Learning Models for One Station

Figure 3 displays the time series and scatter plot of 10-m observed and modelled wind speed over a certain time period for the station Le Havre–Octeville, located at 49.53°N and 0.08°E. The station lies on the coast, in northern France, and is the northernmost station on the Greenwich meridian. This station was chosen as qualitatively representative of the

overall results, and featuring a rather pronounced, but not exceptional, improvement. Other individual stations typically display the same ordering of the performances of the different models, but with rather weaker contrasts for inland stations, and with comparable or greater improvements for many coastal stations. The 10-m wind speed from the ECMWF analyses yields high r.m.s. error (about 2.3 $\mathrm{m\,s^{-1}}$) and low correlation coefficient (about 0.7), as illustrated in the time series (purple line of Fig. 3). The machine-learning models (green and yellow lines in the time series) closely follow the observed wind speed (black line in the time series), suggesting improvements for the r.m.s. error and the correlation coefficient over the ECMWF analyses, as discussed quantitatively below. The scatter plot compares the modelled and observed wind speed for the same time period as that of the time series with the black line indicating perfect representation. The ECMWF model usually overestimates the wind speed over 4 $\mathrm{m\,s^{-1}}$ as can be seen from the scatter plot (represented by purple dots). The implemented models generally underestimate the wind speed over 5 $\mathrm{m\,s^{-1}}$ (illustrated by green and yellow dots). What is the representativity of the high r.m.s. error and the low correlation coefficient for the ECMWF analyses, and of the improved performance by the machine-learning models typical over 8 years for this station?

Figure 4 shows the r.m.s. error and correlation coefficient of all the models (over 8 years) for the reconstruction of the 10-m wind speed at the same station. The r.m.s. error of the ECMWF analyses is high at 2.3 $\mathrm{m\,s^{-1}}$ whereas the correlation coefficient is low at about 0.7. It is conspicuous that all the implemented models imply improvement, resulting in r.m.s. error reduced to values between 1.05 $\mathrm{m\,s^{-1}}$ and 1.35 $\mathrm{m\,s^{-1}}$, and the correlation coefficient increased to values between 0.73 and 0.86. Among the implemented models, one can distinguish three groups. The first group comprises the linear regression models which reduce the r.m.s. error by about 44% and increase the correlation coefficient by about 6%, reducing the inter-quartile range of the r.m.s. error and the correlation coefficient by about 81% and 46% respectively compared to the ECMWF analyses. The second noticeable group is the tree-based machine-learning models which give the best performance, reducing the r.m.s. error by 55% and increasing the correlation coefficient by 22% with a sharp reduction in the inter-quartile range of r.m.s. error and correlation coefficient by 91% and 75% respectively over the ECMWF analyses. The performance of the k-Nearest model is intermediate between the first and the second groups with an improvement in the r.m.s. error and the correlation coefficient by 50% and 15% respectively over the ECMWF analyses.

3.2 Performance of the Parametric and Non-parametric Models over France

In order to obtain a general picture for the whole of France, the above-discussed exercise was carried out for all the stations in France. Figure 5 displays the r.m.s. error and the correlation coefficient of all the models for stations in France. Note that the outliers (defined as values away from the first or third quartile by more than 1.5 times the inter-quartile range) of the ECMWF model have been excluded from the panel for the r.m.s.e. because they were significantly larger than the largest values displayed here.

Overall, all the models generally perform better than the ECMWF analyses in representing the 10-m wind speed (refer also to Tables 5 and 6). Generally, the parametric models (LRall, LRstep, and Lasso) reduce the r.m.s.e. relative to the ECMWF analyses by 25% and increase the correlation coefficient by 8%; all the models reduce the inter-quartile range of the r.m.s. error by approximately 50% and that of the correlation coefficient by 20%. The r.m.s. error of about 25% of the stations in the parametric models are lower than the minimum r.m.s. error represented by the ECMWF model (note that in the figure, the whiskers indicate

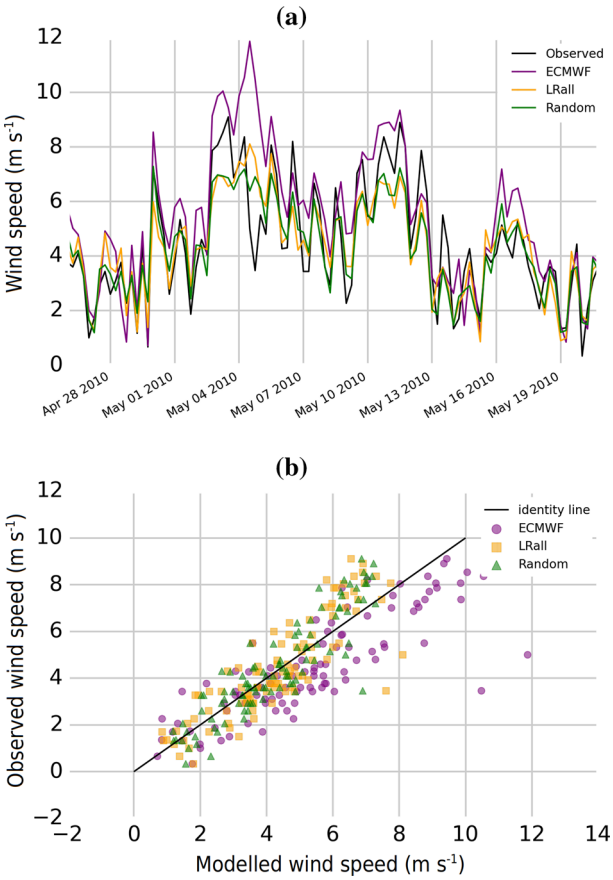

Fig. 3 Time series **a** and scatter plot **b** of the observed and modelled wind speed at 10 m for the Le Havre–Octeville station

the extreme values, but excluding outliers). About 25% of the stations in the ECMWF model display a r.m.s. error higher than the highest value represented by the parametric models. The correlation coefficient of about 50% of the stations in the parametric models are above the third quartile of the ECMWF model.

Overall, the tree-based non-parametric models, such as the Random and Gradient models significantly reduce the r.m.s. error relative to ECMWF analyses by 33% and increase the correlation coefficient by 15%; both reduce the inter-quartile range of the r.m.s. error by roughly 60% and the correlation coefficient by 50%. About 50% of the stations in the tree-based non-parametric models have r.m.s. error lower than the lowest value and correlation coefficient higher than the highest value of the ECMWF model. The r.m.s. error and the correlation coefficient of about 75% of the stations in the Random and Gradient models are well within the first quartile and above the third quartile of the ECMWF model, respectively. Although the performance of the *k*-Nearest model is in between that of parametric and tree-based non-parametric models, there are instances of it degrading the results over the ECMWF analyses. This may be due to the fact that the *k*-Nearest model is sensitive to the number

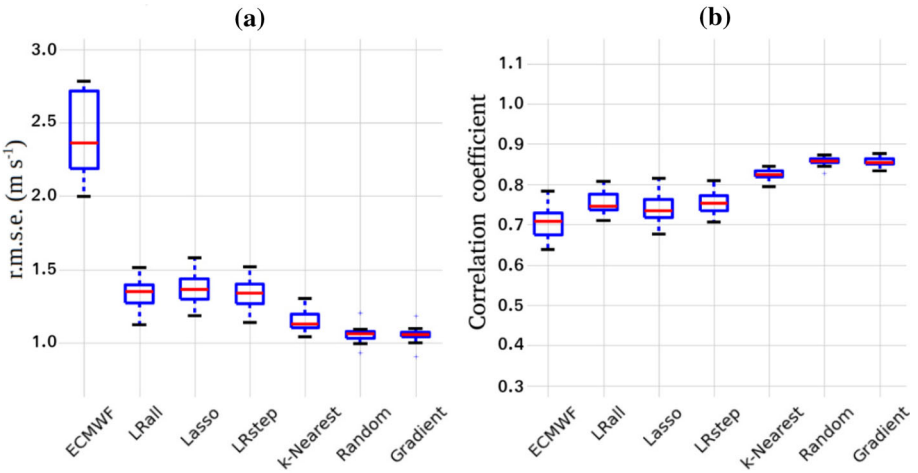

Fig. 4 Root-mean-square error and correlation coefficient of all models for the station Le Havre–Octeville in France. Here and in subsequent plots, the boxplot indicates the median (red), the first and third quartiles (blue box) and the minimum and maximum values (whiskers), excluding outliers (small crosses)

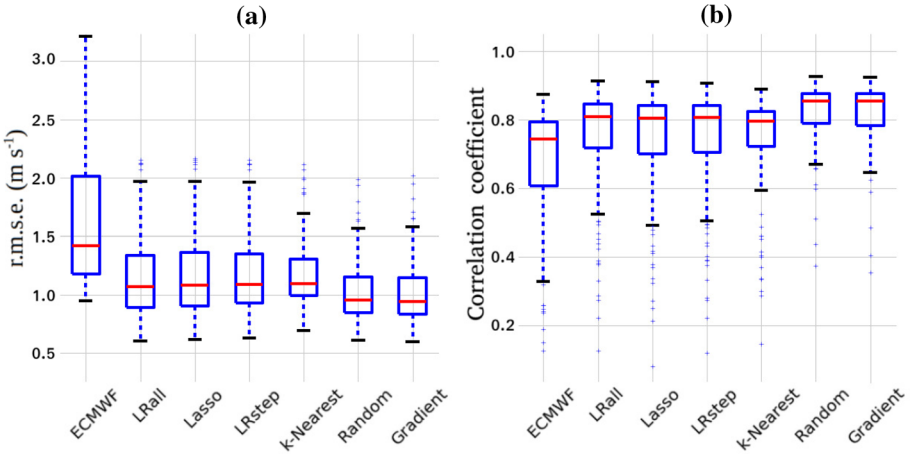

Fig. 5 Root-mean-square error and correlation coefficient of all models for all the stations in France

and kind of variables that are fed and the number k of neighbours chosen. To conclude, the Random and Gradient models seem to provide positive results with minimal effort.

3.3 Geographical Pattern

The improvements obtained by the machine-learning models are not homogeneous geographically. To illustrate this, Fig. 6 shows the percentage change in the r.m.s. error and the correlation coefficient for the LRall model with respect to the ECMWF analyses for stations in France (the geographical patterns for different implementations of random forests are similar between themselves and illustrated in Sect. 4). It is clear that the LRall model performs significantly better than the ECMWF model everywhere. The strongest reductions

Table 5 Quartiles of the r.m.s. error of all models from Fig. 5

Model	Min.	First quartile	Median	Third quartile	Max.	Inter-quartile range
ECMWF	0.94	1.18	1.42	2.02	3.20	0.84
LRall	0.60	0.89	1.07	1.33	1.97	0.44
Lasso	0.62	0.9	1.08	1.36	1.97	0.46
LRstep	0.63	0.93	1.09	1.35	1.96	0.42
k-Nearest	0.69	0.99	1.09	1.30	1.70	0.31
Random	0.60	0.84	0.95	1.15	1.60	0.31
Gradient	0.60	0.83	0.94	1.15	1.62	0.32

Table 6 Quartiles of the correlation coefficients of all models from Fig. 5

Model	Min.	First quartile	Median	Third quartile	Max.	Inter-quartile range
ECMWF	0.32	0.61	0.74	0.79	0.87	0.18
LRall	0.52	0.72	0.81	0.85	0.91	0.13
Lasso	0.49	0.70	0.80	0.84	0.91	0.14
LRstep	0.50	0.70	0.81	0.84	0.91	0.14
k-Nearest	0.60	0.72	0.80	0.82	0.89	0.10
Random	0.68	0.79	0.85	0.88	0.93	0.09
Gradient	0.65	0.78	0.85	0.87	0.92	0.09

in the r.m.s. error of at least 30% are present on the western coast, the southern coast, and Corsica where the ECMWF analyses perform poorly (see Fig. 2). In general, the r.m.s. error of inland stations decreases by 15% with few local stations showing reductions reaching 60%. The correlation coefficients follow a similar pattern with largest increases seen on the coastal stations including Corsica. On average, inland stations show an increase of 6% for the correlation coefficient. The other two parametric models show a pattern similar to the LRall model with the LRstep model performing as well as the LRall model, and the Lasso model performing close to the LRall model (figures for the Lasso and LRstep models are not shown here).

The *k*-Nearest model performs heterogeneously (not shown). The largest reductions of the r.m.s. error are found at the coastal stations including Corsica, whereas a few inland stations yield increases of the r.m.s. error relative to the ECMWF analyses. The mean reduction of the r.m.s. error for the *k*-Nearest model at the coastal stations is larger than that of the parametric models. Yet because of the poor performance of the *k*-Nearest model for the inland stations, parametric models outperform the *k*-Nearest model overall. The results for the correlation coefficient confirm these conclusions.

The performance of tree-based models shows a pattern similar to that of the parametric models but with even larger improvements. The geographical pattern for the performance of the Random model is very similar to the pattern discussed for the Random25 model discussed further (Fig. 9). The changes of the r.m.s. error and of the correlation coefficients relative to the ECMWF model indicate improvement everywhere. The largest changes are found on the western coast, the southern coast, and Corsica with an average reduction of the r.m.s. error of 50% and an increase of the correlation coefficient of 70%. In general, the r.m.s. error of

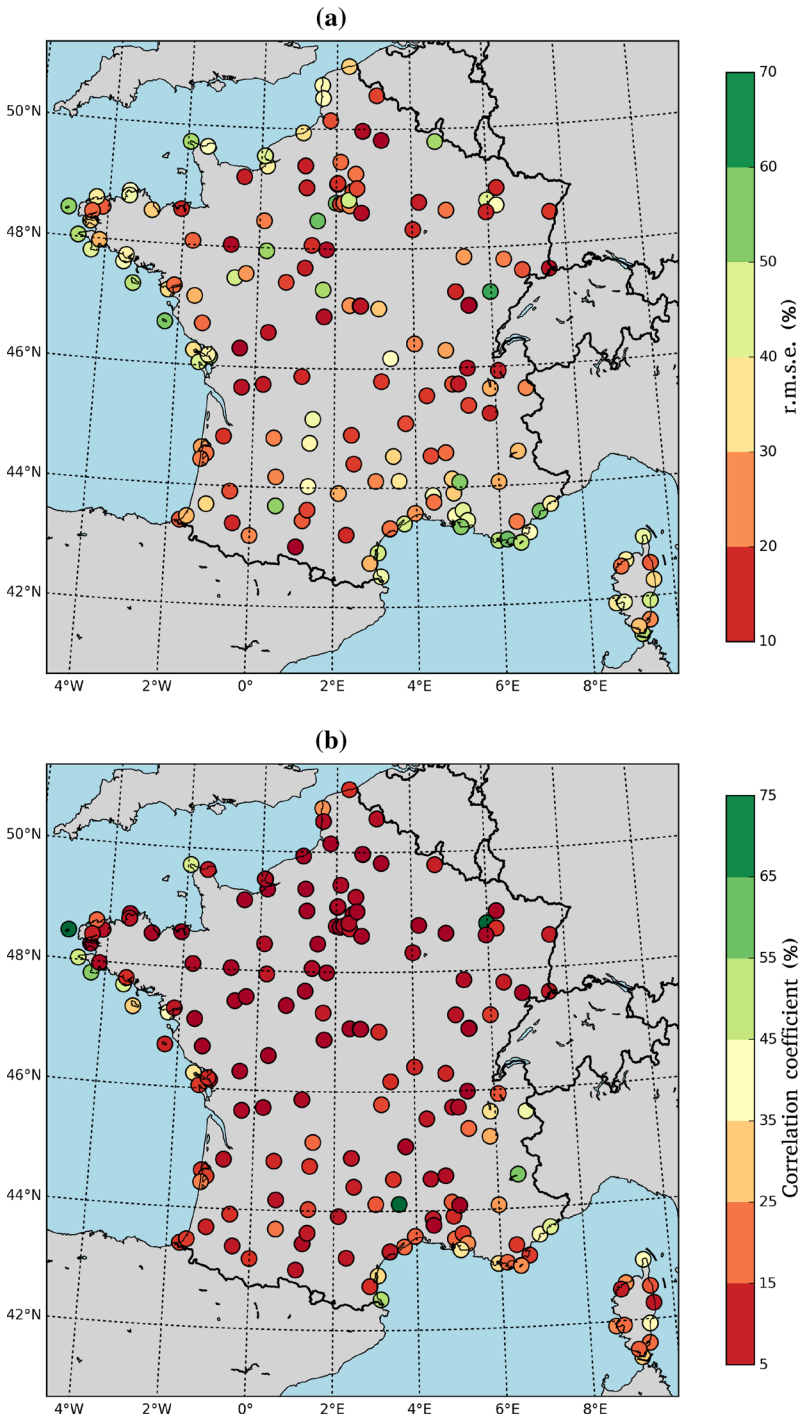

Fig. 6 Percentage change of the r.m.s. error for the LRall model relative to the ECMWF analyses (**a**), and percentage change of the correlation coefficient for the LRall model relative to the ECMWF analyses (**b**)

inland stations decreases by 25% with a few stations showing stronger reductions of up to 60%. The correlation coefficient increases by 12% on average for the inland stations.

As an alternative and complementary diagnostic of the errors obtained by using random forests, the median absolute deviation was also investigated. The median absolute deviation is calculated for each station as the median of the absolute deviation between the modelled wind speed and the observed wind speed. For the estimate of the surface wind speed directly output from the ECMWF analyses, the average over all stations of the median absolute deviation is 0.8 $\mathrm{m\,s^{-1}}$, ranging from 0.5 to 2.4 $\mathrm{m\,s^{-1}}$. For the Random model, the average median absolute deviation is reduced to 0.6 $\mathrm{m\,s^{-1}}$, with a range from 0.35 to 1.3 $\mathrm{m\,s^{-1}}$. The percentage reduction of the median absolute deviation is 26% on average and ranges from 6 to 56%. The geographical patterns for the errors and for the improvements closely resemble those described by Figs. 1 and 9, and hence are not shown.

4 Relevance of the Different Explanatory Variables

The previous section identified the most efficient model and explored the best possible improvements relative to the raw output from the ECMWF analyses. Consequently, we did not restrict the list of explanatory variables: the machine-learning models or selection procedures handled the redundancy or irrelevant information. The inputs for the machine-learning models comprised a long list of explanatory variables which could potentially carry information.

For practical purposes, it is desirable to simplify the implemented models by restricting the list of explanatory variables that carry substantial information. It is instructive to document the list of explanatory variables from which the machine-learning models draw their information.

As the Random model yields the best performance, further investigation is restricted only to applications of the random-forest approach, which provides tools to quantify and rank the relevance of explanatory variables. Our aim is to reduce the list of explanatory variables as much as possible without degrading the performance. To rank and select the most informative explanatory variables, we use the mean decrease in impurity method as described, for instance, in Louppe et al. (2013).

4.1 Reducing the List of Explanatory Variables

In order to develop a simplified and more explainable model, the relevance of explanatory variables for each station in France is analyzed. It is found that the variables relating to velocity dominate the rank table for most of the stations. It is also noted that the ranking of explanatory variables is unique to each station with the importance value in every station dropping typically between the fortieth and the fiftieth variable. This motivated an implementation of the random-forest model with only 50 important explanatory variables specific to each station (compared with 117 explanatory variables for the Random model). This model is named Random50. The performance of the model is not degraded, rather very slightly enhanced; more importantly, it is found that over 50% of the original explanatory variables do not provide useful information. Redundancy in the explanatory variables results from the very high correlation between many explanatory variables. The Random50 model reduces the list of explanatory variables for each station, but in a way specific to each station, and it therefore requires a station-specific analysis. A more generic approach should use the same list of explanatory variables for all the stations. Figure 7 shows the frequency of occurrence

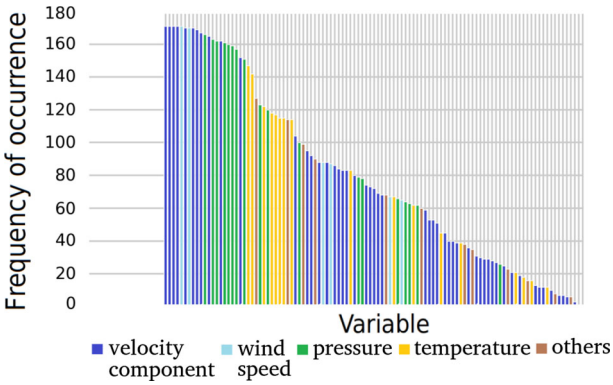

Fig. 7 Frequency of occurrence of explanatory variables for the Random50 model for all the stations in France. Each vertical bar corresponds to one explanatory variable. For readability, we have not included abbreviated names of the variables, but indicate with colours the categories of variables. Note that the explanatory variables based on pressure (to be more precise on the geopotential taken on isobaric surfaces) include the horizontal gradients. These gradients approximate the geostrophic velocity components, hence are closely very related to actual velocity components

of the 50 most important explanatory variables for stations in France. This figure is obtained from the analysis of the lists of 50 most important variables for 171 stations. It should be noted that 107 of the original 117 explanatory variables appear at least for one station's list of 50 most important variables.

A model based on a more generic approach named Random50C with 50 explanatory variables common to all stations was developed and it performs as well as Random50 (Fig. 8). To investigate how much the list of variables can be shortened, another model using random forests, Random25C with the 25 most important explanatory variables was developed. At this point, the performance begins to degrade marginally: the Random25C model performs as well as the Random model, with only a 1% increase of the r.m.s. error overall. However, the Random10C model with the 10 most important variables not only produces an increase of the r.m.s. error by 8% and a decrease of the correlation coefficient by 2%, but it also yields an increase of the inter-quartile range for the r.m.s. error and the correlation coefficient by 13% and 11%, respectively (refer to Tables 7 and 8). Nonetheless, the Random10C model performs significantly better than all the parametric models described in Sect. 3.2.

Further analyzing the list of explanatory variables, we find that the wind speed at 100 m ($F100$), wind difference at 10 m ($F10$) and bulk wind difference between 10 m and 100 m (DF) are the three most significant variables that provide crucial information from the synoptic flow at any given location. Accordingly, another model Random3C with only three variables ($F10$, $F100$ and DF) was set up. The comparison of the performance with those of parametric models turns out to be more nuanced as can be seen from Fig. 8. The performance of a linear regression model with the same three important explanatory variables is poorer than that of the Random3C model (results not shown here). The conclusion of these tests is that a reduction of the explanatory variables to 25 or even to 10 variables is justified and does not significantly affect the performance, but that a reduction to only three explanatory variables is excessive and significantly degrades the performance of the model based on random forests.

Regarding the spatial distribution, the percentage change of the r.m.s. error and the correlation coefficient of the Random25C model with respect to the ECMWF analyses is shown in

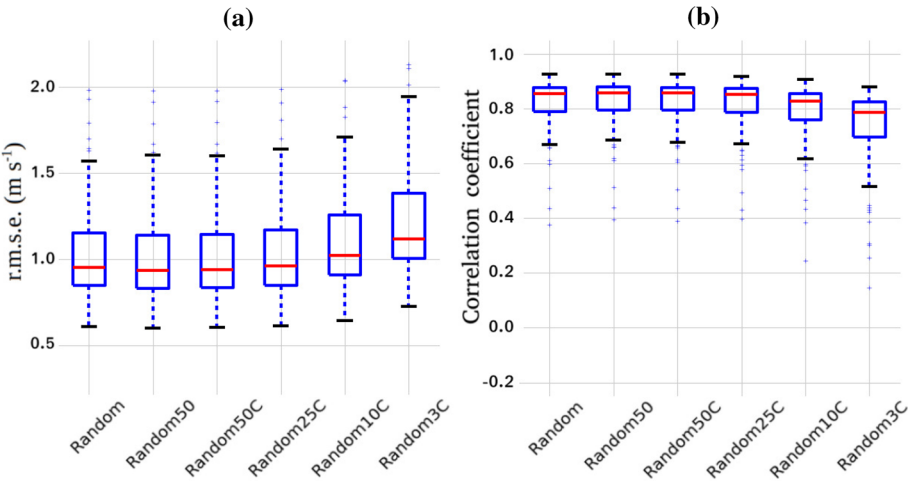

Fig. 8 Root-mean-square error and correlation coefficient of the model using random forests with different lists of explanatory variables for all the stations in France

Table 7 Quartiles of the root-mean-square error of various models using random forests (see Fig. 8), in m s^{-1}

Model	Min.	First quartile	Median	Third quartile	Max.	Inter-quartile range
Random	0.60	0.84	0.95	1.15	1.60	0.31
Random50	0.60	0.83	0.94	1.15	1.63	0.32
Random50C	0.60	0.83	0.94	1.14	1.62	0.31
Random25C	0.61	0.84	0.96	1.16	1.65	0.32
Random10C	0.64	0.91	1.02	1.26	1.70	0.35
Random3C	0.72	1.00	1.12	1.38	1.92	0.38

The extreme values are for the whole dataset, i.e. including outliers

Table 8 Quartiles of the correlation coefficient of various models using random forests (see Fig. 8)

Model	Min.	First quartile	Median	Third quartile	Max.	Inter-quartile range
Random	0.66	0.79	0.85	0.88	0.93	0.09
Random50	0.68	0.80	0.86	0.88	0.93	0.08
Random50C	0.67	0.80	0.86	0.88	0.92	0.08
Random25C	0.66	0.79	0.85	0.87	0.92	0.08
Random10C	0.62	0.76	0.83	0.86	0.91	0.10
Random3C	0.51	0.69	0.79	0.82	0.88	0.13

The extreme values are for the whole dataset, i.e. including outliers

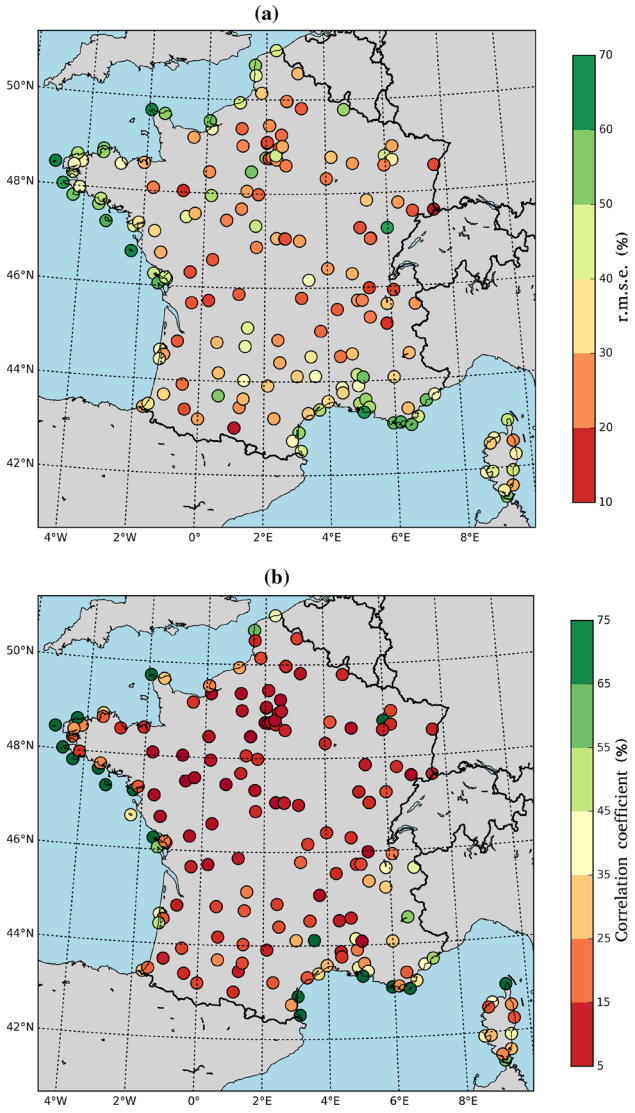

Fig. 9 Percentage change of the r.m.s. error for the Random25C model relative to the ECMWF analyses (**a**), and percentage change in correlation coefficient of the Random25C model relative to the ECMWF analyses (**b**)

Fig. 9. From Fig. 9a it can be seen that the r.m.s. error of inland stations in the north of France is reduced by 30% on average. The r.m.s. error for stations in the inland south decreases on average by 40%. The largest decreases of up to 80% are found for coastal stations in the west, the south and Corsica. Results for the correlation coefficient confirm these conclusions (Fig. 9b). The correlation coefficient for stations in the inland north and inland south show an increase of 15% and 22%, respectively, and for coastal stations the correlation coefficient shows an increase of 60%.

In conclusion, the Random model used an unnecessarily long list of explanatory variables, which is not detrimental to its performance. The Random50C model with 50 common explanatory variables performs as well as the Random50 model but is generic in nature. The Random25C model is simple and robust with just 25 important explanatory variables and is comparable to the Random model in performance. However, with fewer explanatory variables, the Random25C model is not quite as good as the Random50C model. Hence, the Random25C model is a good compromise between performance and simplicity. It is instructive to analyze the list of 25 explanatory variables retained.

4.2 List of Significant Variables

The following are the most significant explanatory variables that bring in unique information to the machine-learning models.
Top 9 list:

- Two horizontal velocity components and the wind speed at 10 m and 100 m above ground (six variables),
- the wind shear between 10 m and 100 m (one variable),
- and the two horizontal components of velocity at 500 hPa (two variables).

Top 25 list (in addition to the previous nine variables):

- the horizontal velocity components at 850 hPa and 925 hPa (four variables),
- the horizontal components of the gradient of geopotential at 925 hPa, 850 hPa and 500 hPa (six variables),
- the horizontal components of the gradients of mean sea level pressure (two variables),
- skin temperature,
- temperature at 2 m,
- the boundary-layer height,
- and one of the components of the gradient of surface pressure.

In the list of the 25 most informative variables, one notes nine variables corresponding to components of the geopotential or pressure. These variables are strongly correlated to the velocity components because geostrophic balance is an excellent approximation at these latitudes. Yet, it is noteworthy that the Random10C model, which does not include these explanatory variables, performed less well than the Random25C model. Finally, it is interesting that few variables describing temperature and boundary-layer parameters appear in the top 25 list. As seen from Fig. 7, variables ranked between 25 and 35 mainly describe the temperature and the boundary layer, yet the comparable performances of the Random25C and the Random50C models suggest that their contribution is minor.

To conclude, it is striking that the most relevant variables almost all describe the flow (velocity components, wind speed, geopotential gradient). It was expected that, given the importance of thermal and convective processes in the boundary layer, the inclusion of information on the temperature and stratification would be helpful. It turns out that this information does not significantly modify the performance of the models. A possible explanation is that the NWP model of the ECMWF already describes rather well the surface flow, and the vertical shear in the first 100 m already encompasses the relevant information on the stratification and mixing in the boundary layer. Another possibility is that we have not provided information on these aspects of the boundary layer with the right choice of explanatory variables.

5 Discussion

This section describes additional work carried out to explore some directions to widen the scope of our results. Indeed, a severe limitation of our approach is that it is only local and it requires prior observations for training the machine-learning models. Hence, it is of great interest to explore and identify patterns in the performance of the models, as this may provide insight regarding the origin of model errors: to what extent does the improvement mainly come from a removal of the bias in the model ouput? Are there errors systematically associated to certain geographical features (mountains, coastlines)? Do the machine-learning models preferentially rely on certain variables in certain geographical contexts? Are there systematic errors associated with other features of the boundary layer (diurnal cycle)?

Regarding the geographical pattern solely based on percentage change of the r.m.s. error and correlation coefficient over the ECMWF model, Fig. 9 gives the impression of three clusters: inland north, inland south, and coastal. We attempt to provide a statistical confirmation in the following section. Another issue is the influence of the time of day on the errors made by machine-learning models, which is addressed in Sect. 5.4.

5.1 Bias

The performance of the machine-learning models is quantified using the r.m.s. error and correlation coefficient as complementary diagnostics. However, it is important to probe how much of the r.m.s. error results from a reduction of a bias present in the ECMWF analyses. For this purpose, the bias of the surface wind speed from the ECMWF analyses is shown in Fig. 10. The locations of the largest r.m.s. error (Fig. 2) coincide with those of the largest bias. There is mostly a positive bias over coastal stations, amounting typically to nearly half of the r.m.s. error. There are also a few inland stations displaying a significant negative bias, associated with unusually large r.m.s. error Over the whole set of stations, the bias is on average 0.5 $\mathrm{m\,s^{-1}}$, amounting to slightly more than a quarter of the r.m.s. error (1.7 $\mathrm{m\,s^{-1}}$). For individual stations, the biases range from -1.6 to 2.5 $\mathrm{m\,s^{-1}}$.

As expected, the machine-learning models prove very efficient at removing the bias. For illustration, Fig. 10 displays the bias for the Random25C model, which is uniformly negligible, the average bias being 0.004 $\mathrm{m\,s^{-1}}$. The bias for individual stations is very weak, ranging for all but two stations from -0.01 to 0.02 $\mathrm{m\,s^{-1}}$. The two outliers are stations in the Alps, with biases between 0.02 and 0.04 $\mathrm{m\,s^{-1}}$.

5.2 Altitude

Topography has a major influence on surface flow, and its description within numerical models remains a challenge. To explore a possible influence on the performance of the models, the relation between percentage change of the r.m.s. error (or of the correlation coefficient) and the altitude of the corresponding station was explored. No evidence for such a link was found. As the local topography has a significant effect on the local variations of the surface flow, we searched for a relation between the small-scale gradient of topography around each station and the model performance. We found no clear link between the gradients of altitude, on a scale of 2 km, and the percentage change of the r.m.s. error and of the correlation coefficient. We further elaborated our previous approach by taking into account the variance of topography around each station. This was achieved by considering the altitude at 0.2 km, 0.5 km, 1 km, 1.5 km, 2 km, 3 km, and 5 km in the north, south, east, and west directions

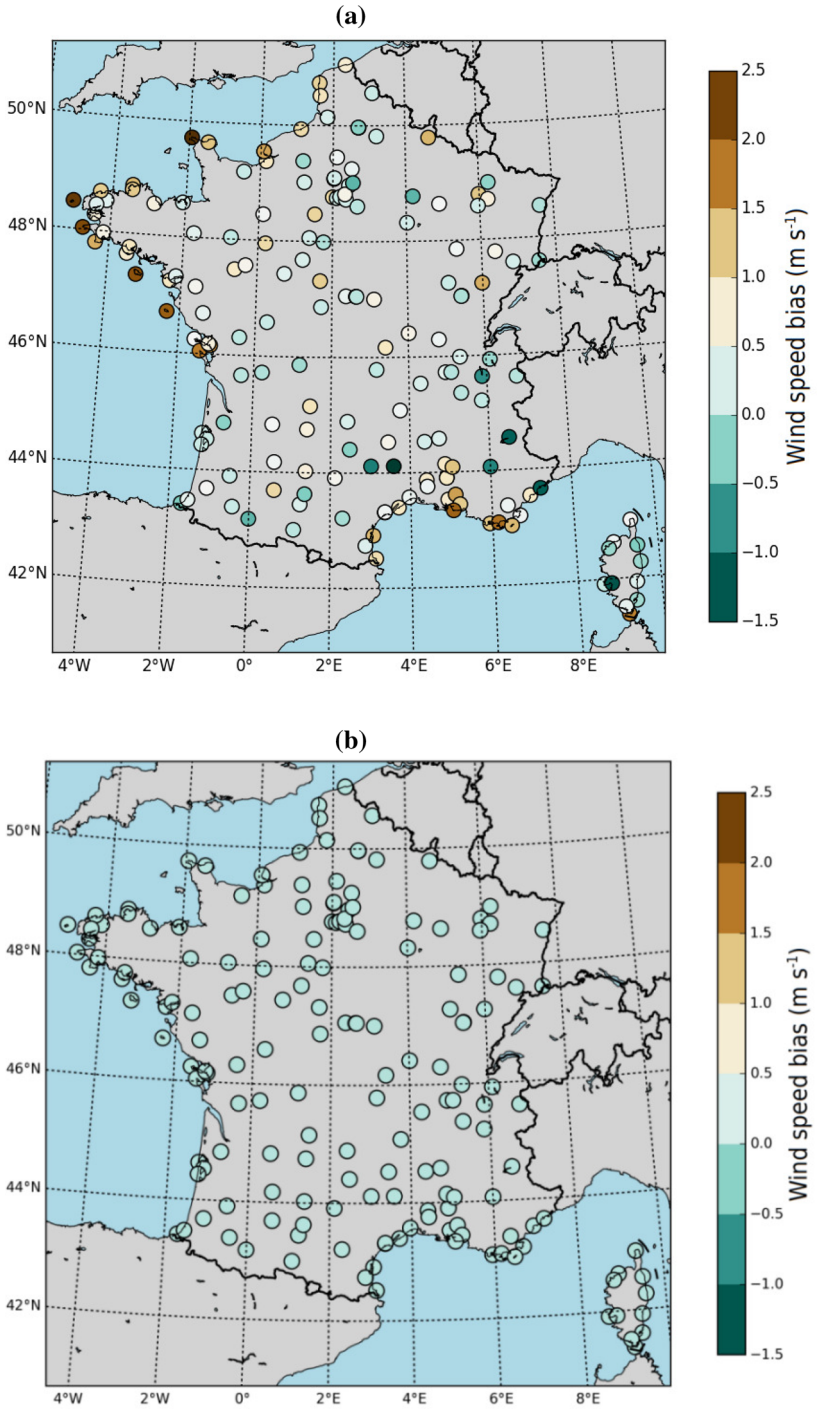

Fig. 10 Bias in the surface wind speed for the ECMWF analyses (**a**) and for the estimated surface wind speed using the Random25C model (**b**)

around each station and by calculating the overall variance of altitude. No clear link between the variations of the r.m.s. error and correlation coefficient with the altitude parameters was discovered.

5.3 Clustering

Independently, unsupervised classification using k-means clustering was also performed by using the r.m.s. error and the correlation coefficient from the ECMWF analyses, and percentage change in the r.m.s. error and the correlation coefficient of the Random25C over the ECMWF analyses as explanatory variables. This approach to identify clusters did not yield conclusive results. More work is needed to identify clusters which could provide insights on the origins of the errors in the NWP model used.

A different interrogation also relating to the spatial organisation of errors is on the spatial correlation between errors. Useful information can be obtained from the spatial correlation of model errors (Livezey and Chen 1983; Elmore et al. 2006; Wilks 2006). In our case, the correlation of time series of errors at different stations was investigated for a few test stations, and revealed that the correlation coefficients were always low (smaller than 0.1) and with no apparent spatial structure (not shown). The contrary (large-scale structure in the correlations) would have certainly indicated that we were not using properly the information on the large-scale state of the atmosphere from the NWP model.

5.4 Diurnal Cycle

Inspection of the error at specific stations suggests that a diurnal cycle for the error could be present. This is in part natural, as there is a marked diurnal cycle in the properties of the boundary layer (thermal mostly, but also, to a lesser degree, involving wind speed). To illustrate this diurnal cycle, the probability density functions of errors for the four different analysis times (0000, 0600, 1200, and 1800 UTC) are shown in Fig. 11 at Le Havre–Octeville station and for the ECMWF analyses. Biases are clearly present and vary with the time of day. The signs of these biases are not robust across stations, and should not be judged as representative. Attempts were made to remedy this diurnal cycle by training four different machine-learning models, one for each time of day. This procedure provided only mild and inconclusive improvement, and hence is not documented here. The purpose of this paragraph is rather to highlight this issue for further exploration, and for which a better knowledge of the modelling system and its limitations may be particularly beneficial. Sensitivity of the errors to season was also investigated for a few stations, but this preliminary exploration did not suggest a conspicuous signal, so it was not pursued further.

5.5 Application to Other Variables

The methodology described was applied to 10-m wind speed because of the demand from the wind-energy industry for better estimates for surface wind speed. The methodology is not specific to wind speed, however, and could apply to other quantities. For the velocity components, it has been applied in preliminary tests as an intermediate step before calculating the wind speed. This did not produce improved performance for the end calculation of the wind speed and was not pursued. The methodology could also be used with wind direction, for which statistical estimators can also be used despite its cyclic character (e.g. Yamartino 1982).

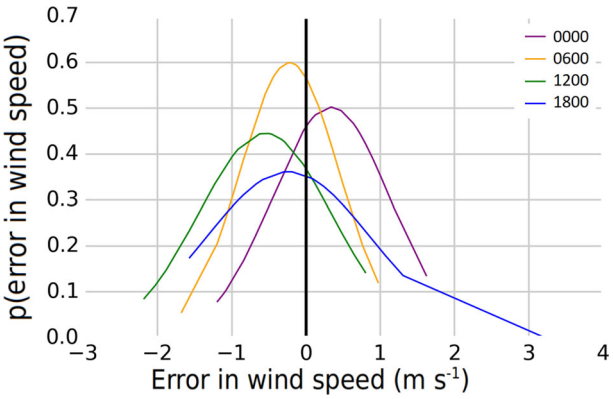

Fig. 11 Probability density function of the error for the wind speed at 10 m from the Random25C model at various hour indices for the station Le Havre–Octeville

Variables different from the wind speed, notably temperature, could be estimated with the above methodology. However, the errors of NWP models for temperature are less of a concern than for the surface flow. The r.m.s. error and correlation coefficient of the temperature at 2 m directly output from the ECMWF analyses was calculated for all the 171 stations (not shown). The average r.m.s. error was 1.5 K with a standard deviation of 0.8 K, indicating strong variations among stations. Indeed, for individual stations, the r.m.s. error ranges from 0.8 to 8.0 K. Excluding four stations that appear as outliers brings the average r.m.s. error down to 1.3 K, with a standard deviation of 0.4 K. The average correlation coefficient is 1.0, the weakest correlation being 0.9. Given the good performances of the direct model output, the possible relative gain from statistical post-processing is weaker.

6 Exploratory Test Using Forecasts

We have explored the relationship between outputs of a NWP model and the observed wind speed at 10 m from 171 stations in France. We have shown that post-processing using machine-learning models could provide significant improvements over the performance of the numerical weather prediction model alone. Before reporting our conclusions in Sect. 7, we need to consider an essential question hitherto left aside: the NWP outputs were extracted from analyses. In practice, it is forecasts that are of use for wind-energy operators. Does the relationship identified between model outputs and observed wind speed hold when the explanatory variables are taken from forecasts? Are the improvements from machine-learning models applied to forecasts comparable to those obtained from analyses? Below, we probe this issue for the case of one station, encouragingly suggesting that our results carry over fully to forecasts.

This section emphasizes the improvement of the forecasts of the surface wind speed from the outputs of the ECMWF model, using the same post-processing as described in previous sections. Note that this provides only a lower bound on the potential accuracy of forecasts, because the machine-learning models are not trained on the forecasts and do not use all the available information (see discussion below).

The ECMWF high-resolution global-forecast model is run twice a day at a base time of 0000 and 1200 UTC and each run forecasts the weather up to 10 days. We limit this study

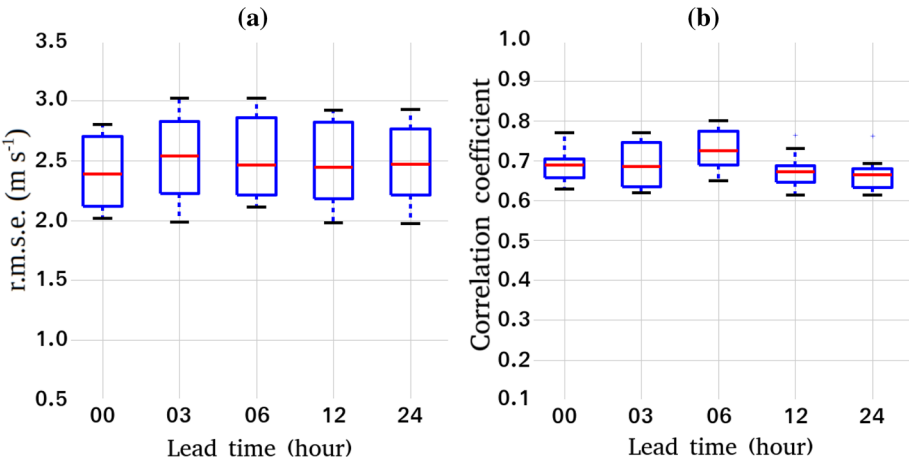

Fig. 12 Root-mean-square error and correlation coefficient of the ECMWF forecast wind speed at 10 m at various time horizons for the station Le Havre–Octeville

to the station Le Havre–Octeville (already used in Sect. 3.1). Appropriately, the ECMWF forecast data were retrieved at lead times of 0 h, 3 h, 6 h, 12 h, and 24 h, where 0 h corresponds to that of the analyses. The machine-learning models used to reconstruct the wind speed from these forecasts are the same as described and used previously, i.e. they have been trained using model outputs from the analyses. In other words, there has not been a new machine-learning model trained with outputs from the forecasts.

To describe the baseline, Fig. 12 shows the r.m.s. error and the correlation coefficient of the wind speed at 10 m from ECMWF forecasts at various lead times for the station Le Havre–Octeville in France. As seen previously, the r.m.s. error is rather large (nearly 2.5 $\mathrm{m\,s^{-1}}$), and it remains fairly constant over the first 24 h of the forecast.

The Random25C model is trained on the analyses as described in Sect. 4, and applied to the outputs of the ECMWF forecasts at lead times from 3 to 24 h. The r.m.s. error and correlation coefficient of the reconstructed wind speed are shown in Fig. 13. Strikingly, the r.m.s. error is dramatically reduced (down to 1.2 $\mathrm{m\,s^{-1}}$ or less, with a very narrow spread): the average reduction in r.m.s. error over all the lead times is about 55%, and the average increase in correlation coefficient is about 21%. These improvements are simply consistent with those obtained with random forests from the analyses (Sect. 4). There is a suggestion of a slight time evolution of the accuracy, with a maximum accuracy for lead times of 6 h, which could be explored if the investigation at other stations confirmed it to be a robust feature. The message to retain here is that the improvements carry over to forecasts, and that for lead times up to 24 h, these improvements are fairly stable in time. Hence, this approach holds promise for forecasting. The results could be further improved by applying a model that is trained separately for each lead time directly on the forecasts. This, and the investigation over all stations in France, are topics for future research.

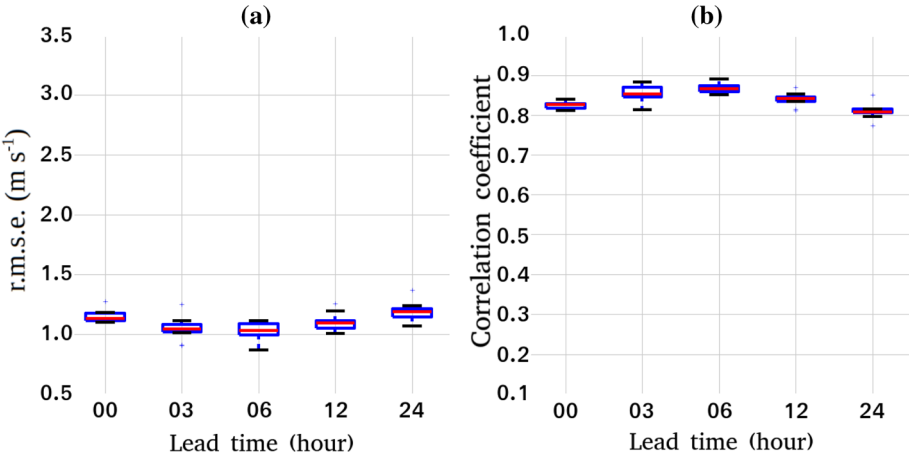

Fig. 13 Root-mean-square error and correlation coefficient of wind speed at 10 m from the Random25C model at various time horizons for the station Le Havre–Octeville

7 Conclusion

Several parametric and non-parametric machine-learning models are used to estimate the 10-m wind speed from the analyses of the ECMWF model at 171 stations in France. Two issues were particularly emphasized: first, the use and comparison of both parametric models (multi-linear regression, as in a majority of model output statistical practices) and machine-learning methods (notably random forests), and second, the identification of model variables that carried information for the estimation of the surface wind speed.

The ECMWF model estimates well the wind speed at 10 m in the inland north of France. However, there are significant errors in the wind-speed estimation on the coasts, the inland south and Corsica. The mean r.m.s. error and correlation coefficient of all the stations in France from 2010 to 2017 are 1.74 m s^{-1} and 0.68, respectively. For machine-learning models, as explanatory variables, we retained model variables describing velocity, geopotential, and temperature at several levels, along with their vertical and horizontal gradients. We also included certain variables describing the boundary layer.

All the machine-learning models, parametric and non-parametric, generally bring an improvement in the estimation of the wind speed relative to the direct output of the ECMWF model, as intended. All the parametric models (linear regression) show a similar performance with an average decrease of 25% of the r.m.s. error and an increase of 8% of the correlation coefficient. Tree-based non-parametric models (random forest and gradient boosting) show the best performance with a mean decrease of 33% of the r.m.s. error and an increase of 15% of the correlation coefficient. The *k*-Nearest model, being not only non-parametric, but also data sensitive, gave intermediate results. The largest improvements in performance by all the models are found on the coastal stations on the North Sea and the Atlantic coast, on the Mediterranean coast and in Corsica.

The contribution of various explanatory variables in capturing the relationship between synoptic circulations and local flows has been investigated. The random-forest machine-learning approach is simple and robust, requiring almost no data preparation, and it also provides tools to quantify and rank the relevance of explanatory variables. The random-forest model with 50 explanatory variables common to all stations has the best performance

in terms of objective scores. Curtailing the list of explanatory variables to 25 simplifies the model and only marginally degrades the performance. Further reducing the list of explanatory variables noticeably degrades the results (see Tables 7 and 8; for instance, the medians of the r.m.s. error for models Random50C, Random25C, Random10C and Random3C are, respectively, 0.94, 0.96, 1.02 and 1.12 $\mathrm{m\,s^{-1}}$). Hence, the random-forest model with 25 variables common to all stations (Random25C) appears to be the best compromise between performance and simplicity. A generic list of 25 most significant variables to estimate the wind speed at any location was proposed. It is striking to note that the most relevant variables are almost exclusively flow variables (velocity components, wind speed or components of the gradient of geopotential). Revisiting this with particular care to provide better information on the stratification near the surface (e.g., through an estimation of a bulk Richardson number) would be worthwhile to make this more conclusive.

Further issues, such as the geographical pattern of model performance or its dependence upon local topography, have been explored. Upon looking at the figures showing the percentage improvement in r.m.s. error and correlation correlation, there seems to appear a geographical pattern (with the largest improvements on the coast and the inland south, and moderate improvements in the inland north). Preliminary attempts to objectively define geographical clusters of stations showing similar model performance were hampered by outliers, and more research would be needed in this direction. Attempts to test the sensitivity of the machine-learning models to local topography (altitude, its gradients or small-scale variance) did not reveal any conspicuous relationship. Finally, the presence of a diurnal cycle in the bias of the ECMWF model was detected in certain stations. A preliminary attempt was carried out to remedy this, but it was too limited in time and concerned only one station so it remained inconclusive. This aspect requires further, more systematic investigation.

This study confirms, for the estimation of surface wind speed, the relevance of machine-learning models, such as random forests, in agreement with the findings and choices of Zamo et al. (2016). These authors, in the context of providing improved, gridded data of surface flow, used random forests and explored strategies for obtaining gridded surface velocity components over a whole territory, not just at a given location where observations have been available. Our results on the comparison of parametric and non-parametric models, on the geographical distribution of improvements, and on the relevance and selection of explanatory variables are complementary. The very encouraging test using forecasts in Sect. 6 opens the way for further studies to apply these models for forecasts, notably for wind energy, estimating wind speed at 100 m. Further improvements and enhanced interpretations of the machine-learning models could be explored using approaches described recently (McGovern et al. 2019). Another important source of information to tap into are outputs from numerical weather prediction models at higher resolution. The French meteorological agency, Météo-France, produces forecasts for mainland France at a higher spatial resolution (horizontal grid-spacing of 1.3 km presently). Investigating the performance of machine-learning models using input from higher resolution models constitutes a topic for further research.

Acknowledgements This research was supported by the Agence Nationale pour la Recherche (ANR), through the Project FOREWER (ANR-14-CE05-0028).

References

Alonzo B, Plougonven R, Mougeot M, Fischer A, Dupré A, Drobinski P (2018) Forecasting and risk management for renewable energy, Springer, chap From Numerical Weather Prediction outputs to accurate local surface wind speed: statistical modeling and forecasts

Alpaydin E (2010) An introduction to machine learning, 2nd edn. Massachusetts Institute of Technology Press, Cambridge

Baars J, Mass C (2005) Performance of National Weather Service forecasts compared to operational, consensus and wieghted model output statistics. Weather Forecast 20:1034–1047. https://doi.org/10.1175/WAF896.1

Bauer P, Thorpe A, Brunet G (2015) The quiet revolution of numerical weather prediction. Nature 525:47–55. https://doi.org/10.1038/nature14956

Breiman L (2001) Random Forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Chang WY (2014) A literature review of wind forecasting methods. J Power Energy Eng 2:161–168

de Rooy W, Kok K (2004) A combined physical-statistical approach for the downscaling of wind speed. Weather Forecast 19:485–495

Devis A, van Lipzig N, Demuzere M (2013) A new statistical approach to downscale wind speed distributions at a site in northern Europe. J Geophys Res Atmos 25:2272–2283. https://doi.org/10.1002/jgrd.50245

Elmore K, Baldwin M, Schultz D (2006) Field significance revisited: spatial bias errors in forecasts as applied to the ETA model. Mon Wea Rep 134:519–531

Foley A, Leahy P, Marvuglia A, McKeogh E (2012) Current methods and advances in forecasting of wind power generation. Renew Energy 37:1–8. https://doi.org/10.1016/j.renene.2011.05.033

Gareth J, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning. Springer, Berlin

Glahn H, Lowry D (1972) The use of model output statistics (MOS) in objective weather forecasting. J App Meteor 11:1203–1211

Haiden T, Janousek M, , Bidlot JR, Buizza R, Ferranti L, Prates F, Vitart F (2018) Evaluation of ecmwf forecasts, including the 2018 upgrade. ECMWF Technical Memo 831, 10.21957/ldw15ckqi

Horlacher V, Osborne S, Price J (2012) Comparison of two closely located meteorological measurement sites and consequences for their areal representativity. Boundary-Layer Meteorol 142:469–493. https://doi.org/10.1007/s10546-011-9684-3

Kalnay E (2003) Atmospheric modeling, data assimilation and predictability. Cambridge University Press, Cambridge

Kang JH, Suh MS, Hong KO, Kim C (2011) Development of updateable model output statistics (UMOS) system for air temperature over South Korea. Asia-Pac J Atmos Sci 47:199–211. https://doi.org/10.1007/s13143-011-0009-8

Lazic L, Pejanovic G, Zivkovic M, Ilic L (2014) Improved wind forecasts for wind power generation using the Eta model and MOS (Model Output Statistics). Energy 73:567–574

Livezey R, Chen W (1983) Statistical field significance and its determination by Monte Carlo techniques. Mon Weather Rev 111:46–59

Louppe G, Wehenkel L, Sutera A, Geurts P (2013) Understanding variable importances in forests of randomized trees. In: Advances in Neural Information Processing Systems, 26

McGovern A, Lagerquist R, Jergensen G, Elmore K, Homeyer C, Smith T (2019) Making the black box more transparent. Bull Am Meteor Soc. https://doi.org/10.1175/BAMS-D-18-0195.1

Mejia J, Giordano M, Wilcox E (2018) Conditional summertime day-ahead solar irradiance forecast. Sol Energy 163:610–622

Muller JP, Massaron L (2016) Machine learning for dummies. Wiley, Hoboken

Ranaboldo M, Giebel G, Codina B (2013) Implementation of a model output statistics based on a meteorological variable screening for short-term wind power forecasts. Wind Energy 16:811–826

Riddaway B (2013) Newsletter no. 136 - summer 2013 https://www.ecmwf.int/node/14593

Salameh T, Drobinski P, Vrac M, Naveau P (2009) Statistical downscaling of near-surface wind over complex terrain in southern France. Meteorol Atmos Phys 103:253–265. https://doi.org/10.1007/s00703-008-0330-7

Schmeits M, Kok K, Vodelezang D (2005) Probabilistic forecasting of (severe) thunderstorms in the Netherlands using model output statistics. Wea Forecast 20:134–148

Schuhen N, Thorarinsdottir T, Gneiting T (2012) Ensemble model sutput statistics for wind vectors. Mon Weather Rev 140:3204–3219

Smith A, Lott N, Vose R (2011) The Integrated Surface Database: Recent Developments and Partnerships. Bull Am Meteor Soc 92:704–708. https://doi.org/10.1175/2011BAMS3015.1

Tascikaraoglu A, Uzunoglu M (2014) A review of combined approaches for prediction of short-term wind speed and power. Ren Sust Energy Rev 34:243–254. https://doi.org/10.1016/j.rser.2014.03.033

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Roy Stat Soc Series B 58:267–288
Wang X, Guo P, Huang X (2011) A review of wind power forecasting models. Energy Procedia 12:770–778. https://doi.org/10.1016/j.egypro.2011.10.103
Wilby R, Dawson C (2013) The Statistical Downscaling Model: insights from one decade of application. Int J Climatol 33:1707–1719. https://doi.org/10.1002/joc.3544
Wilks D (2006) On "field significance" and the false discovery rate. J Appl Meteor 45:1181–1189
Wilson L, Vallée M (2002) The Canadian updateable model output statistics (UMOS) system: design and development tests. Weather Forecast 17:206–222
Yamartino R (1982) A comparison of several 'single-pass' estimators of the standard deviation of wind direction. J Clim App Met 23:1362–1366
Zamo M, Mestre O, Arbogast P, Pannecouke O (2014) A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production, part I: Deterministic forecast of hourly production. Sol Energy 105:792–803. https://doi.org/10.1016/j.solener.2013.12.006
Zamo M, Bel L, Mestre O, Stein J (2016) Improved gridded wind speed forecasts by statistical postprocessing of numerical models with block regression. Weather Forecast 31:1929–1945. https://doi.org/10.1175/WAF-D-16-0052.1