**Noname manuscript No.**
(will be inserted by the editor)

# Using machine learning methods to improve surface wind from the outputs of a Numerical Weather Prediction model

**Naveen Goutham · Bastien Alonzo · Aurore Dupré · Riwal Plougonven · Rebeca Doctors · Lishan Liao · Mathilde Mougeot · Aurélie Fischer · Philippe Drobinski**

Received: DD Month YEAR / Accepted: DD Month YEAR

**Abstract** The relation between the outputs of a Numerical Weather Prediction (NWP) model and the observed surface winds is explored using statistical and machine learning models. Eight years of wind measurements at a height of 10 m (from 2010 to 2017) from 171 stations spread over mainland France and Corsica are used as reference. Operational analyses from the European Center for Medium Range Weather Forecasts (ECMWF) provide the model information not only on the surface wind, but on other aspects of the atmospheric state at the location (or aloft of) 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 ECMWF modelled wind, by itself, has Root Mean Square Errors (RMSE) over all stations distributed widely around a median of 1.42 m $s^{-1}$. Using statistical post-processing and making use of a a historical set of data for training, the median of the RMSE at all stations can be reduced down to 1.07 m $s^{-1}$ with linear regressions, and down to 0.94 m $s^{-1}$ with random forests or gradient boosting. Enhanced improvements are found for coastal stations, where the errors were largest. Random forests are further explored to trim down the list of explanatory variables: a list of 25 explanatory variables, mainly consisting of wind variables (wind, horizontal gradients of geopotential on different isobaric surfaces, shear between 10 and 100 m) and marginally including 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 at a given time is used on forecasts of the NWP model, for lead times up to 24 hours. The statistical/machine learning model is found 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 forecasts.

**Keywords** Downscaling · Machine learning · Surface wind

## 1 Introduction

Surface winds are a meteorological variable of considerable importance because they impact human activities in a number of ways, including damage to buildings, fallen tower cranes, and injuries due to

N. Goutham, B. Alonzo, A. Dupré, R. Plougonven, R. Doctors, L. Liao and Ph. Drobinski
Laboratoire de Météorologie Dynamique/IPSL, Ecole Polytechnique, CNRS,
Palaiseau, France
E-mail: riwal.plougonven@lmd.polytechnique.fr

· M. Mougeot
Department of Applied Mathematics, Ecole Nationale Supérieure d'Informatique pour l'Industrie et l'Entreprise,
Evry, France.

A. Fischer
Laboratoire de Probabilités et Modèles Aléatoires,
Université Paris Diderot, Paris 7, Paris, France.

objects carried in strong winds. Over the past decade, rapid development of wind energy has created a new motivation and demand for estimations of winds near the surface. Notably, the evolution of regulations for the pricing of wind energy (from feed in tariffs to market prices) imply an increased demand for accurate forecasts of surface winds at wind farm locations.

Numerical Weather Prediction (NWP) models constitute a major source of information on surface winds. However, as surface winds are turbulent and strongly influenced by small-scale features absent in the limited representation of NWP models, the modelled surface winds, when compared to local observations at a given site, generally exhibit large errors, including biases. Now, for a given site where observations are available for a long enough interval, it is logical to try and use these observations to learn from and correct the model's biases and errors for that location. In fact, estimating a local quantity from output of a NWP model and past observations at a given location has been an active field of research for half a century, generally called Model Output Statistics (MOS, [GL72]). [GL72] have applied multilinear regressions to several variables, including surface wind, using a forward stepwise screening procedure to select the variables used as predictors. Nowadays, it is common for operational centers to carry out MOS to provide forecasts of quantities where observations are available [WV02, BM05, SKV05, KSHK11, ZMAP14]. As weather forecasts evolve in nature, from deterministic to probabilistic, some of the approaches used for MOS have also evolved [STG12].

Fundamentally, the endeavour to estimate a small-scale, unresolved, fluctuating quantity from modelled knowledge of the large-scale field connects to several research fields with different aims, different sources of information, and different criteria for validation. One is MOS, stated above, which generally focuses on a given location for which observations are available. Another name is *downscaling*, i.e. building 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 (SDSM, [WD13]). Downscaling applied to surface winds has been applied to estimate surface winds with an emphasis on identifying variables which carry information [SDVN09, DvLD13]. For locations in Southern France, where topographic effects crucially affect the winds, [SDVN09] used generalized additive models to estimate wind components from outputs of the ERA-Interim reanalysis from the European Centre for Medium-Range Weather Forecasts (ECMWF).

Finally, the need to estimate sub-grid scale components of the flow from modelled knowledge of the large-scale flow motivate the development of parametrizations in weather and climate models (e.g. [Kal03]). 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, donscaling studies inform us on the portion 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 is in the scope of MOS or downscaling, i.e. improving the estimation of surface winds, 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 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 [dRK04]. The sources of errors can be identified as:

- model error: the model describes the atmospheric flow only approximately, partly because of discretization and limited resolution, partly because processes that occur on small scales are represented through parametrizations.
- representativity error: the model value represents some average over space. For a variable like surface wind 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., [HOP12]).
- predictability limits (when considering forecasts): even if the model is perfect, errors, however small in the initial states, will 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 [Kal03].

The skill of NWP models is continuously increasing [BTB15], as are their spatial resolutions. Both elements imply that the models' description of surface winds is improving. Surface winds as they are directly output from NWP models still suffer from significant errors [HJ$^+$18]. Other variables, in particular large-scale variables like pressure, will be more accurate.

The question of precisely estimating winds at specific locations has received recently renewed interest from the wind energy sector. Very different approaches have been considered for forecasting wind 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 have been developed using a variety of techniques (eg. [Cha14], [TU14], [FLMM12], [WGH11]). For longer lead times, from half a day to several days, output from NWP models have been used, including MOS approaches for wind speed [RGC13, LPZI14] and for solar irradiance [MGW18]. 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. [RGC13] present a stepwise screening procedure to identify the most relevant variables to forecast surface winds at two locations, showing that variables describing the wind lead to the best performances.

The purpose of the present study is to explore and improve the estimation of 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, emphasis is put on evaluating the improvement, for the estimation of the surface winds, coming from machine learning models. Another objective is to try and identify those variables in the NWP model output that carry the most information to reconstruct the surface winds.

The present study builds on the exploration of parametric and non-parametric models for surface winds introduced in [APM$^+$18]. In that study, one specific location was considered, allowing a detailed exploration of regression models at that particular site. It was found that the best performance was obtained with linear regression, considering appropriate variables. Random forests performed nearly as well, without the need for a detailed expertise. The present study extends 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. It leads us to understand how the performance varies from one geographical area to another.

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

## 2 Data and Methodology

### 2.1 Data

The Integrated Surface Database (ISD) is a global database of observed weather data available at 1-hour frequency [SLV11]. About 400 weather stations in France update their weather data on ISD. ISD-Lite is a subset database of hourly time series of original data with fewer variables and in an easy-to-use format specifically made available for research activities. In order to better train the models, we decided to work on stations with over 90% of available data for a span of 8 years, 2010-2017. As a result, we retrieved observed data from 171 stations well distributed across mainland France and Corsica.

The ECMWF is an intergovernmental operational center that provides medium-range weather forecasts on a global scale. It has the largest repository of archived global weather data. ECMWF operational analyses[1] are retrieved with a spatial resolution of 0.125° in latitude and longitude over

[1] best 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

mainland France and Corsica. While this is a fine resolution for global NWP output, this remains coarse-grained when comparing surface wind to measurements at one specific location, given for instance the sensitivity to the local topography.

The local surface wind is related to the synoptic-scale flow in the atmosphere. The large-scale (synoptic) systems like depressions, fronts, and storms are described in terms of physical variables at different pressure levels such as 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 temperature, heat flux, surface pressure, and 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 vertical wind shear and the temperature gradient. Information about those was computed from the raw data as shown in Table 3.

**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 wind component | $m.s^{-1}$ | $u$ |
| 1000/925/850/500 | meridional wind 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 among the NWP model's surface variables. The last three variables are accumulated over the last six hours.

| **Altitude** | **Variable** | **Unit** | **Symbol** |
|---|---|---|---|
| $10m/100m$ | wind speed | $m.s^{-1}$ | F |
| $10m/100m$ | zonal wind component | $m.s^{-1}$ | u |
| $10m/100m$ | meridional wind component | $m.s^{-1}$ | v |
| $2m$ | temperature | $K$ | t2m |
| surface | skin temperature | $K$ | skt |
| msl | mean sea level pressure | $Pa$ | msl |
| surface | surface pressure | $Pa$ | sp |
| - | boundary layer height | m | blh |
| - | boundary layer dissipation | $J.m^{-2}$ | bld |
| surface | surface latent heat flux | $J.m^{-2}$ | slhf |
| surface | surface sensible heat flux | $J.m^{-2}$ | sshf |

**Table 3** Explanatory variables computed as differences in the vertical between two heihgt or pressure levels.

| **Vertical level** | **Variable** | **Unit** | **Symbol** |
|---|---|---|---|
| $10m$ to $100m$ | bulk wind shear | $m.s^{-1}$ | DF |
| $1000hPa$ to $925hPa$ | bulk wind shear | $m.s^{-1}$ | DFP |
| $1000hPa$ to $925hPa$ | temperature difference | $K$ | DTP |

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 4 closest ECMWF grid points surrounding that station. We also computed additional set of quantities by taking north-south (NS), east-west (EW), and diagonal gradients around each station, estimated using finite differences. We observed that the north-south and east-west gradients were found to be 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 (NS and EW) 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-hour frequency of the ECMWF model outputs, we defined a 2-hour *averaging window* by only considering the observed data at the hour, an hour before and after the top of the hour, at 00H, 06H, 12H and 18H.

The ability of the ECMWF model to represent the observed wind speed is quantified by the Root Mean Square Error (RMSE) denoted by $E_{w,obs}$, and Pearson's correlation $\rho_{w,obs}$, given in Eqs. (1) and (2) respectively. Here, $w$ stands for the ECMWF time series 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 $|A|$ standing for the number of elements of a set $A$, and $\bar{y} = \frac{1}{|\mathcal{S}|}\sum_{t\in\mathcal{S}} y_t$ is the mean of the time series $y$.

Figure 2 shows the RMSE and correlation between observed and 10 m wind speed from the ECMWF analysis for the meteorological stations under consideration in France. Figure 2a shows that the RMSE of ECMWF exceeds 1.0 $m.s^{-1}$ for most of the inland stations: the minimum at an individual station is 0.95 $m.s^{-1}$, the maximum is 4.58 $m.s^{-1}$. The average over all stations is 1.74 $m.s^{-1}$, with a standard deviation of 0.79 $m.s^{-1}$. The coastal stations in the west, south and Corsica have a higher RMSE of at least $2m.s^{-1}$. In Figure 2b, we see that the correlation 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, due to higher RMSE and lower correlation observed along the coasts, special attention was paid to these stations during interpolation to check if the location of ECMWF grid points has an effect. Upon careful examination, it was noticed that the location of grid points have no significant influence. This degradation may be due to factors that likely contribute to the difficulty of modeling surface wind 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 sea breeze.

We computed year by year the average RMSE and correlation of the ECMWF analyses over all stations (see Figure 1). An increase of the performance of the model in the year 2014 is observed (RMSE decrease) resulting from changes in the ECMWF model which affected surface winds, notably a modification of the parametrization of surface drag and the upgrade of the vertical resolution, going from 91 to 137 levels in June 2013 [Rid13]. Nevertheless, the upgrade did not have an impact on the correlation. The average RMSE and correlation for the time period 2010–2017 are 1.74 $m.s^{-1}$ and 0.68 respectively. It is also instructive to include the median of the RMSE for all stations: 1.42 $m.s^{-1}$. This value is smaller than the average, which is expected for a positive variable such as RMSE, which can be very large in locations where the model performs very poorly. These errors are significant given that the time-averaged wind averaged over all stations is 3.4 $m.s^{-1}$. More precisely we calculated the ratio of RMSE to the time-averaged wind for each station. The overall mean of these ratios is 0.52, implying that any significant decrease of the error is worthwhile.

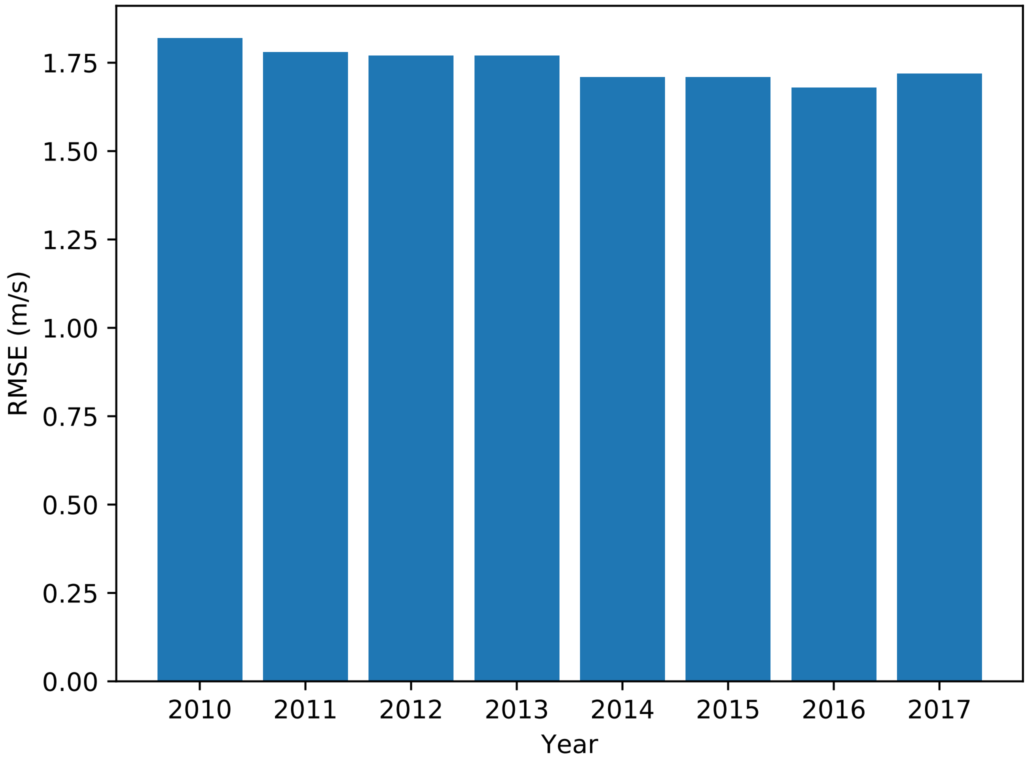



**Fig. 1** RMSE of the raw output of the ECMWF analyses for the 10 m wind speed over all the stations in France for the years 2010-2017. Extreme values are 1.82 (in 2010) and 1.68 $m.s^{-1}$ (in 2016), and the average is 1.74 $m.s^{-1}$. The correlation is stable during this period and equal to 0.68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

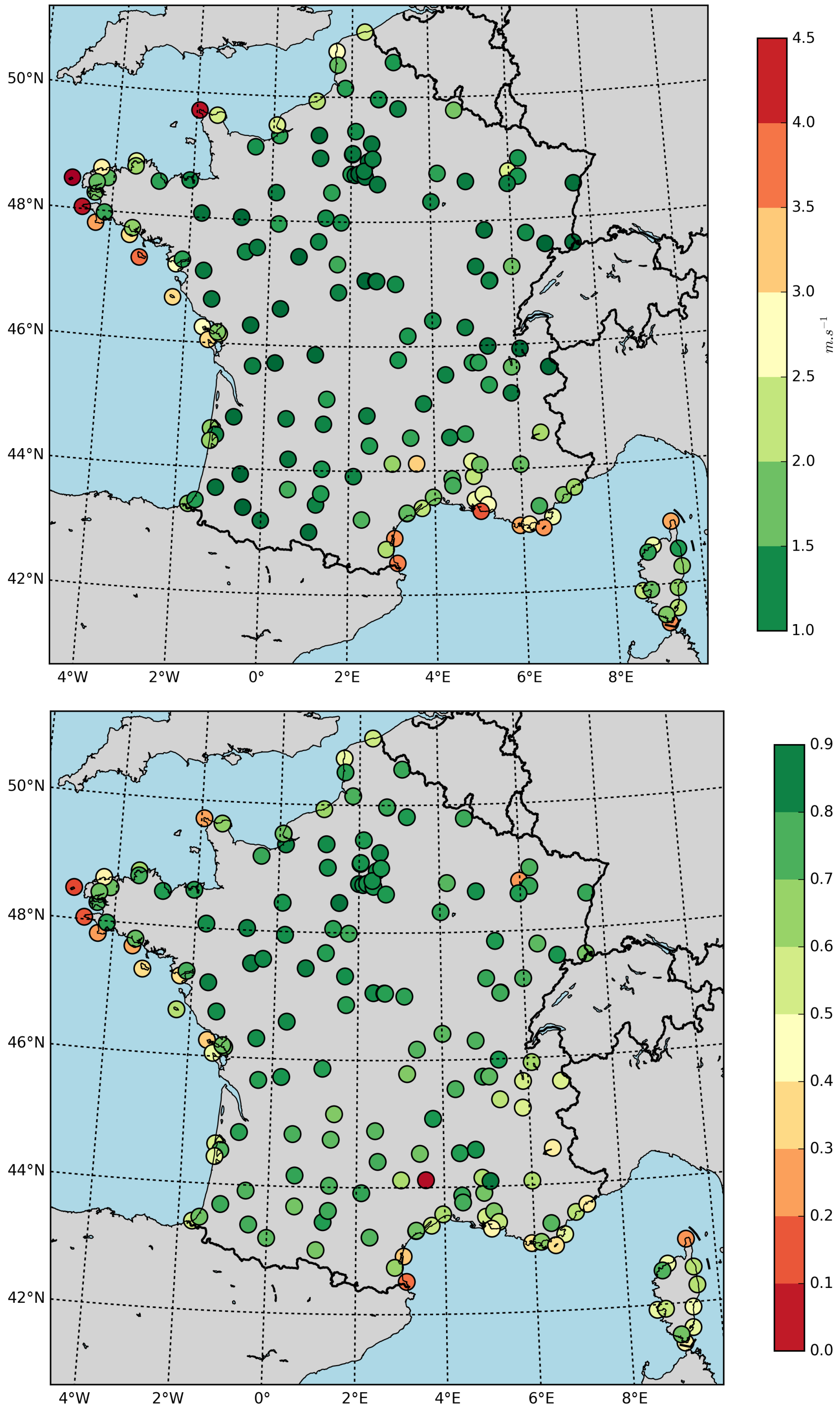



**Fig. 2** RMSE of 10 m wind speed in the *ECMWF* analysis (top panel); correlation of 10 m wind speed in the *ECMWF* analysis (bottom panel). Here and in following boxplot figures, standard definitions are used: the red bar indicates the median, the box is delimited by the first and third quartiles. The whiskers indicate the minimum and maximum values, aside from outliers, which are identified as less than the first quartile, or greater than the third quartile, by more than 1.5 times the interquartile range.

## 2.2 Methodology

The aim of this work is to model observed wind speed at the above mentioned meteorological stations in France from the outputs of the ECMWF model, starting from the study in [APM+18]. Here, the target variable is the observed wind speed. The aim is to model this wind speed using as explanatory variables only output from 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 output and inputs may be described analytically, based on some probability distribution (for instance, Gaussian model). On the contrary, a non-parametric method does not rely on a particular distribution assumption on the data, but involves several tuning parameters.

### *2.2.1 Linear Regression*

*Linear regression* is a widely used model, which tries to identify a linear relationship between the response $Y_t$ and the explanatory variables $X_t^1, X_t^2, ...X_t^p$ at time $t$:

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

where the $\beta_j$s are the regression coefficients that need to be estimated using least-square approach, and $\varepsilon_t$ 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 forward or backward, to retain only a subset of the explanatory variables. Forward selection starts with an empty list of predictors adding one highly significant predictor at each step until a stopping criterion is reached, whereas backward selection starts with a full list of predictors eliminating one highly insignificant predictor at each step until a stopping criterion is reached. Without Gaussian assumption, *Lasso regression* (also called $\ell^1$ regularization) may be employed to select the most important predictors by adding a penalty term to the least-square error. This penalty acts as a constraint favoring a weaker sum of the absolute values of the regression coefficients; this leads some of the coefficients to shrink to zero, implying that the corresponding explanatory variable is dropped [GWHT13, Tib96].

### *2.2.2 Random Forests*

In non-parametric frameworks, decision trees are today commonly used for modeling. Decision trees split iteratively the input space by minimizing the target variance on each side of the split [MM16]. Decision trees have the advantage of being easy to set up and understand, and can capture non-linear 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) is used: it consists in drawing multiple subsets for training the model (bootstrap), and then aggregating together the resulting trees. The variance is correspondingly lower, the risk of overfitting is much reduced. This method has been demonstrated to significantly enhance accuracy. It can be further improved by an additional modification in the procedure, and this leads to *random forests*. 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 data set. 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. In other words, it consists in 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 [GWHT13], they are based on decision trees which are fairly easy to understand, and they 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 Neighbor* algorithm. It takes the $k$ closest training observations based on Euclidean distance and predicts the output as the average of the $k$ nearest neighbors outputs. 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 is a procedure to define split the data into a *training* dataset, and a dataset to *test* and evaluate the perfomance of the model. The data set is partitioned into 10 subsets. Training is performed in a cyclic way on 9 subsets keeping the last one to evaluate the model. Global performances are computed by averaging the 10 repetitions. The Python packages used in the work are NumPy, SciPy, matplotlib, pandas, and Scikit-Learn.

## 3 Comparison of Different Parametric and Non-Parametric Models

The performance of the machine learning (ML) models relative to ECMWF raw model output is computed using a relative error for both RMSE and correlation as follow:

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

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

where $E_{ECMWF}$ and $\rho_{ECMWF}$ are the RMSE and correlation computed between the ECMWF model and the observation; $E_{ML}$ and $\rho_{ML}$ are respectively the RMSE and correlation computed between the predictions of a given machine learning or statistical model and the observations.

The parametric models implemented in this work are *Linear Regression with all explanatory variables* (hereafter $LR_A$), *Linear Regression with stepwise selection of variables* (hereafter $LR_{SW}$), and *the Ordinary Least Square (OLS) with lasso regularization* (hereafter $LR_{\ell_1}$). The non-parametric models implemented in this work are *Random Forest with all variables* (hereafter $RF_A$), *Gradient Boosting* (hereafter $GB$), and *k-Nearest Neighbors* (hereafter $KNN$) using the 10 most important explanatory variables provided by the Random Forest model.

All models are summarized in the following Table 4. Two more $KNN$ models were also trained but are not mentioned in this paper because of their poor performances: one with all explanatory variables and another with only 5 wind related explanatory variables.

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

| machine learning model | Name |
|---|---|
| Linear Regression with all variables | $LR_A$ |
| Linear Regression with stepwise selection | $LR_{SW}$ |
| OLS with lasso regularization | $LR_{L1}$ |
| Random Forest with all variables | $RF_A$ |
| Gradient boosting with all variables | $GB$ |
| K-nearest neighbor with the best 10 chosen variable from $RF_A$ | $KNN$ |

### 3.1 Performance of the machine learning models for one station

Figure 3 illustrates the time series and scatter plot of 10 m observed and modeled winds over a certain time period for the station Le Havre-Octeville located (49.53° N and 0.08° E): it lies on the coastline, in northern France, and is the northernmost station on the Greenwhich meridian, on the northern bank of the Seine estuary. This station was chosen as qualitatively representative of the overall results, but featuring a rather pronounced, but not exceptionnal, 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 foor many coastal stations. The 10 m wind speed from the $ECMWF$ analysis has high RMSE (about 2.3 $m\,s^{-1}$) and low correlation (about 0.7), as illustrated in the time series (purple line of figure 3). The machine learning models (green and yellow lines in the time series) closely follow the observed wind (black line in the time series), suggesting improvements in RMSE and correlation over $ECMWF$, as discussed quantitatively below. The scatter plot shows the modeled and observed winds plotted against each other for the same time period as that of the time series with the black line indicating perfect representation. The $ECMWF$ model usually overestimates winds over 4 $m.s^{-1}$ as can be seen from the scatter plot (represented by purple dots). The implemented models generally underestimate winds over 5 $m.s^{-1}$ (illustrated by green and yellow dots). Is the representativity of high RMSE and low correlation by $ECMWF$, and improved performance by the machine learning models typical over 8 years for this station?

Figure 4 shows the boxplot of RMSE and correlation of all the models (over 8 years) for the reconstruction of 10 m wind speed at the same station. It can be seen that the RMSE of $ECMWF$ is high at 2.3 $m.s^{-1}$ whereas the correlation is low at about 0.7. It is conspicuous that all the implemented models bring in improvement, with RMSE reduced to values between 1.05 $m.s^{-1}$ and 1.35 $m.s^{-1}$, and correlation increased to values between 0.73 and 0.86. Among the implemented models, one can distinguish three groups based on improvement over $ECMWF$. The first group is the machine learning models which improve RMSE by about 44% and correlation by about 6%, bringing down the Inter Quartile Range (IQR) of RMSE and correlation by about 81% and 46% respectively compared to $ECMWF$. The second noticeable group is that of the tree based machine learning models which give the best performance: they reduce RMSE by 55% and increase correlation by 22% with a sharp reduction in the IQR of RMSE and correlation by 91% and 75% respectively over $ECMWF$. The performance of the $KNN$ model is intermediate between the first and the second groups with an improvement in RMSE and correlation by 50% and 15% respectively over $ECMWF$.

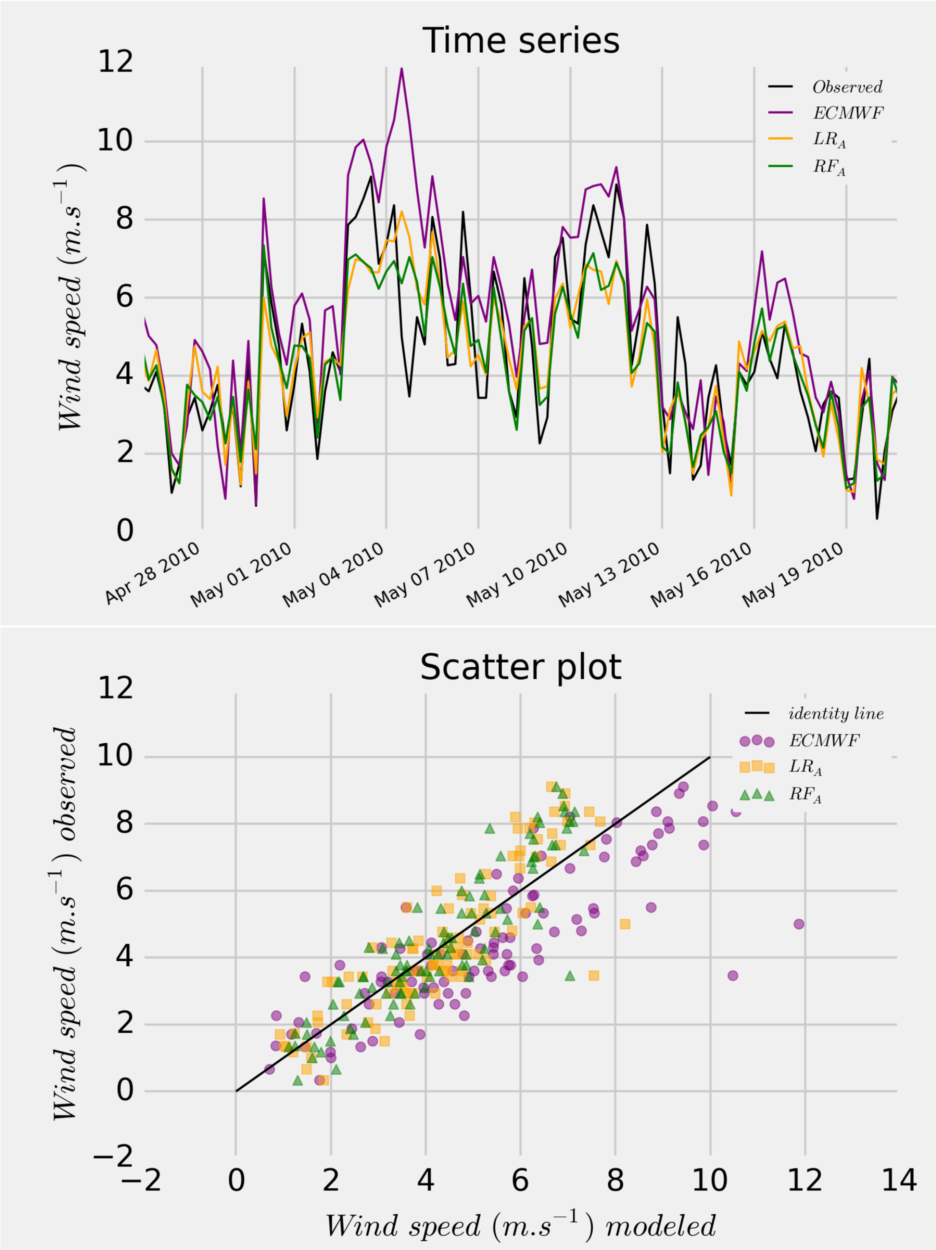



**Fig. 3** Time series (top) and scatter plot (bottom) of the 10 m observed and modeled winds for the Le Havre-Octeville station

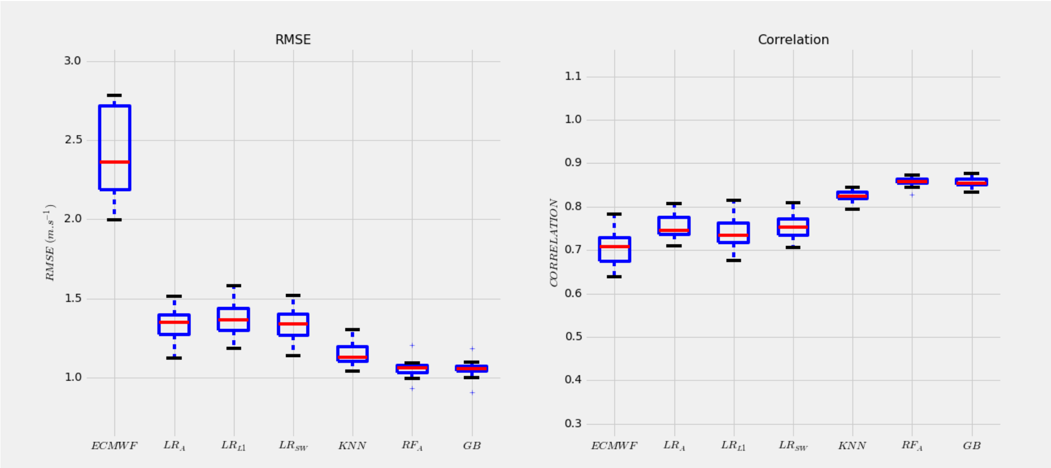



**Fig. 4** Boxplot of RMSE and correlation of all models for the station Le Havre-Octeville in France

After having seen the results for one station, how do the general results look like if the same exercise is done on all the stations in France? This question will be answered in the following section.

## 3.2 Performance of the parametric and non-parametric models over France

In order to have a general picture for the whole of France, the above discussed exercise was reproduced for all the stations in France. Figure 5 shows the boxplot of RMSE and correlation of all the models for stations in France. Note that the outliers (defined using the standard defintion, i.e. further of the first or third quartile by more than 1.5 times the inter-quartile range) of the $ECMWF$ model have been excluded from the RMSE boxplot as they were significantly higher and distorting the scale of the plot.

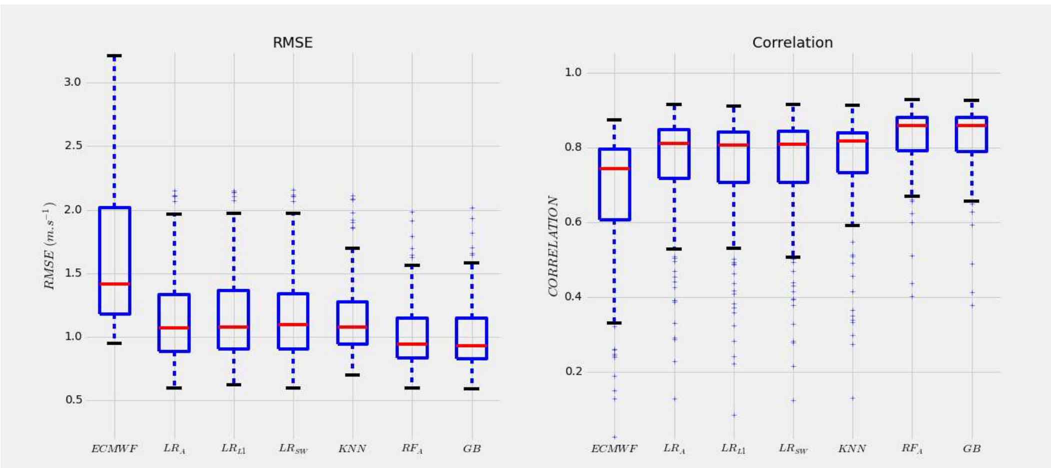



**Fig. 5** Boxplot of RMSE and correlation of all models for all the stations in France

Overall, it can be observed that all the models generally perform better than $ECMWF$ in representing 10 m wind (refer also to Table 5 and Table 6). Generally, the parametric models ($LR_A$, $LR_{SW}$, and $LR_{L1}$) improve the RMSE over $ECMWF$ by 25% and correlation by 8%; all of them reducing the IQR of RMSE by approximately 50% and correlation by 20%. The RMSE of about 25% of the stations in the parametric models are below the minimum RMSE represented by the $ECMWF$ model (note that in the boxplots, the whiskers indicate the extreme values, but excluding outliers, see

Fig. 2). About 25% of the stations in the $ECMWF$ model have RMSE higher than the highest value represented by the parametric models. The correlation of about 50% of the stations in the parametric models are above the third quartile (Q3) of the $ECMWF$ model.

Overall, the tree based non-parametric models such as $RF_A$ and $GB$ significantly improve the RMSE over ECMWF by 33% and correlation by 15%; both of them reducing the IQR of RMSE by roughly 60% and correlation by 50%. About 50% of the stations in the tree based non-parametric models have RMSE lower than the lowest value and correlation higher than the highest value of the $ECMWF$ model. The RMSE and correlation of about 75% of the stations in the $RF_A$ and $GB$ models are well within the first quartile (Q1) and above the third quartile (Q3) of the $ECMWF$ model respectively. Although the performance of the $KNN$ model is in between that of parametric and tree based non-parametric models, there are instances of it degrading the results over $ECMWF$. This may be due to the fact that the $KNN$ model is sensitive to the number and kind of variables that are fed and the number of k-neighbors chosen. To conclude, $RF_A$ and $GB$ models seem to provide robust results with minimal efforts.

**Table 5** Quartiles of the RMSE of all models from the boxplot (Figure 5)

| Model | Min | Q1 | Median | Q3 | Max | IQR |
|---|---|---|---|---|---|---|
| $ECMWF$ | 0.94 | 1.18 | 1.42 | 2.02 | 3.20 | 0.84 |
| $LR_A$ | 0.60 | 0.89 | 1.07 | 1.33 | 1.97 | 0.44 |
| $LR_{L1}$ | 0.62 | 0.9 | 1.08 | 1.36 | 1.97 | 0.46 |
| $LR_{SW}$ | 0.63 | 0.93 | 1.09 | 1.35 | 1.96 | 0.42 |
| $KNN$ | 0.69 | 0.99 | 1.09 | 1.30 | 1.70 | 0.31 |
| $RF_A$ | 0.60 | 0.84 | 0.95 | 1.15 | 1.60 | 0.31 |
| $GB$ | 0.60 | 0.83 | 0.94 | 1.15 | 1.62 | 0.32 |

**Table 6** Quartiles of the correlation of all models from the boxplot (Figure 5)

| Model | Min | Q1 | Median | Q3 | Max | IQR |
|---|---|---|---|---|---|---|
| $ECMWF$ | 0.32 | 0.61 | 0.74 | 0.79 | 0.87 | 0.18 |
| $LR_A$ | 0.52 | 0.72 | 0.81 | 0.85 | 0.91 | 0.13 |
| $LR_{L1}$ | 0.49 | 0.70 | 0.80 | 0.84 | 0.91 | 0.14 |
| $LR_{SW}$ | 0.50 | 0.70 | 0.81 | 0.84 | 0.91 | 0.14 |
| $KNN$ | 0.60 | 0.72 | 0.80 | 0.82 | 0.89 | 0.10 |
| $RF_A$ | 0.68 | 0.79 | 0.85 | 0.88 | 0.93 | 0.09 |
| $GB$ | 0.65 | 0.78 | 0.85 | 0.87 | 0.92 | 0.09 |

### 3.3 Geographical Pattern

The improvements obtained by the machine learning models are not homogeneous geographically. To illustrate this, Figure 6 shows the percentage change in RMSE and correlation of $LR_A$ model with respect to $ECMWF$ 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 $LR_A$ model improves the RMSE and correlation over the $ECMWF$ model everywhere. Greatest improvements in RMSE of at least 30% could be noticed on the Western coast, the Southern coast, and Corsica where $ECMWF$ had performed poorly (ref Figure 2). In general, the RMSE of inland stations improves by 15% on average with few local stations showing higher improvements of up to 60%. Correlation follows a similar pattern with highest improvements seen on the coastal stations including Corsica. On an average, inland stations show an improvement of 6% in correlation. The other two parametric models show a pattern similar to $LR_A$ model with $LR_{SW}$ performing as good as $LR_A$, and $LR_{L1}$ performing close to $LR_A$ (figures of $LR_{L1}$ and $LR_{SW}$ not shown here).

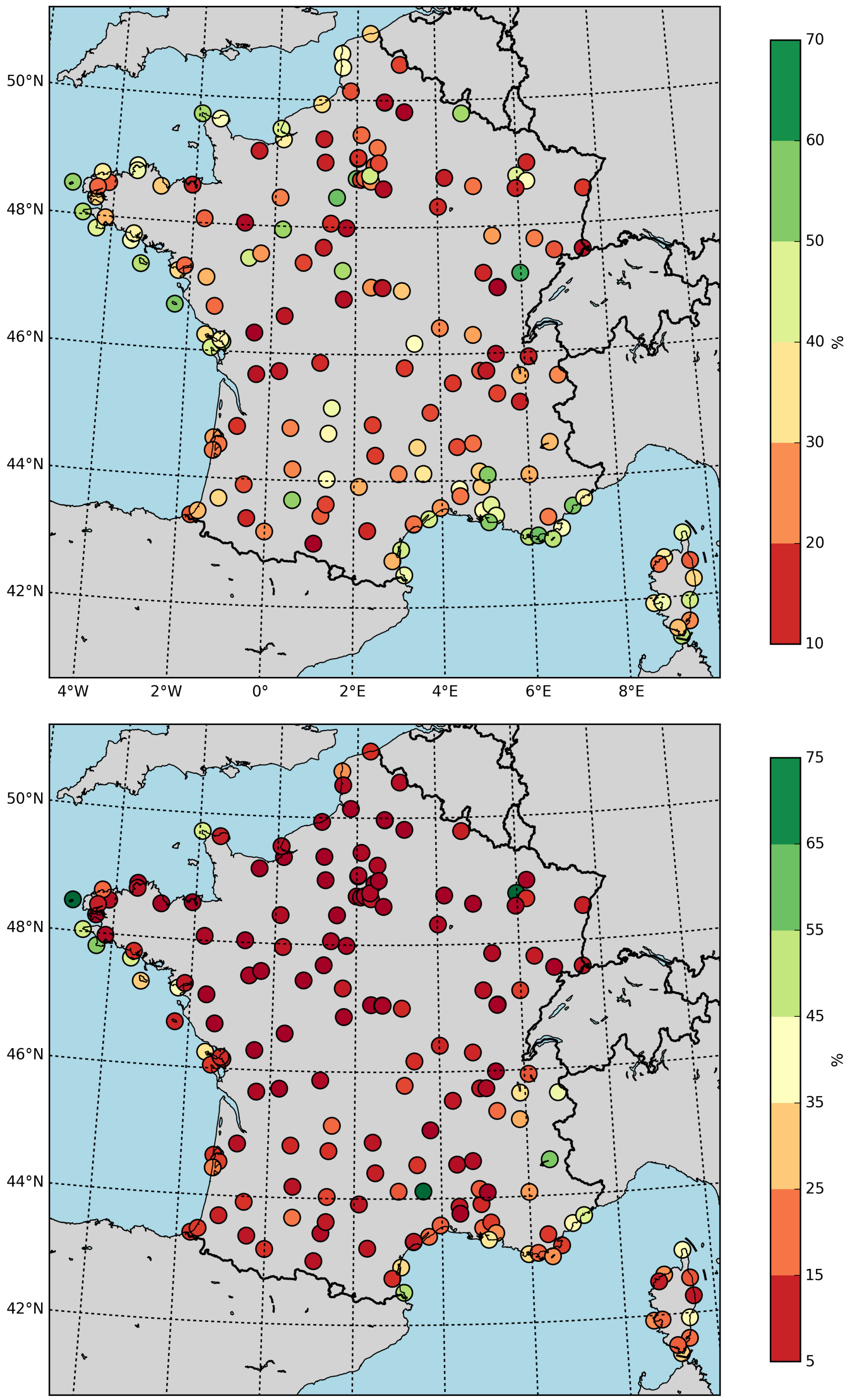



**Fig. 6** Percentage change in RMSE of $LR_A$ model with respect to $ECMWF$ analyses (top), and percentage change in correlation of $LR_A$ model with respect to $ECMWF$ analyses (bottom).

The $KNN$ model has mixed performance (figures not enclosed here). The highest improvements in RMSE could be observed at the coastal stations including Corsica, whereas few inland stations suffer degradation in RMSE over $ECMWF$. The mean improvement in RMSE of KNN model at the coastal stations is higher than that of the parametric models. As a result of general degradation of RMSE in the inland stations, parametric models outperform $KNN$. More degradation than improvement could be noticed when it comes to the correlation of KNN model. Although there is an improvement in correlation at the coastal stations compared to the parametric models, the inland stations suffer significant degradation.

The tree-based models show a a pattern similar to parametric models but with even higher improvements. The geographical pattern for the performance of the $RF_A$ model is very similar to the pattern discussed for the $RF_{C25}$ discussed further below (Figure 9). These models also improve the RMSE and correlation over the $ECMWF$ model everywhere. Greatest performance could be seen on the Western coast, the Southern coast, and Corsica with an average improvement in RMSE of 50% and correlation of 70%. In general, the RMSE of inland stations improves by 25% on average with few local stations showing higher improvements of up to 60%. Correlation shows an improvement of 12% on average in the inland stations.

As $RF_A$ model is simple and robust providing the best performance; it will be further explored in the section that follows.

## 4 Relevance of the Different Explanatory Variables

The aim of the previous section was to identify the most efficient model and to explore the best possible improvements relative to the raw output from $ECMWF$. Consequently, we did not restrict the list of explanatory variables (letting the machine learning models or selection procedures handle the redundancy or irrelevant information). We fed the machine learning models with 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 only to those that carry substantial information. It will also be instructive to document the list of explanatory variables from which the machine learning models draw their information.

As $RF_A$ yielded the best performance, the further work will be restricted only to the $Random\ forest$ model. Moreover it provides tools to quantify and rank the relevance of explanatory variables. Our aim will be to reduce the list of explanatory variables as much as possible without degrading the performance.

### 4.1 Reducing the List of Explanatory Variables

With an objective to develop a simplified and a more explainable model, the relevance of explanatory variables for each station in France was analyzed. It was observed that the wind variables dominated the rank table in most of the stations. It was also noted that the ranking of explanatory variables was unique to each station with the importance value in every station dropping typically between the $40^{th}$ and the $50^{th}$ variable. This led to try another $Random\ forest$ model $RF_B$ with only 50 important explanatory variables specific to each station (compared to the p=117 number of initial variables). The performance of the model was not degraded, rather very slightly enhanced; more importantly, it was found that over 50% of the original explanatory variables were not providing useful, additional information. A redundancy in the explanatory variables as a result of very high correlation between them was observed. The $RF_B$ model reduced the list of explanatory variables for each station, but in a way specific to each station, and therefore requiring 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 of the 50 most important explanatory variables for stations in France. This list was developed by grouping the list of 50 most important variables for 171 stations. It should be noted that 107 of the original 117 explanatory variables appear in the 50 most important variables list.

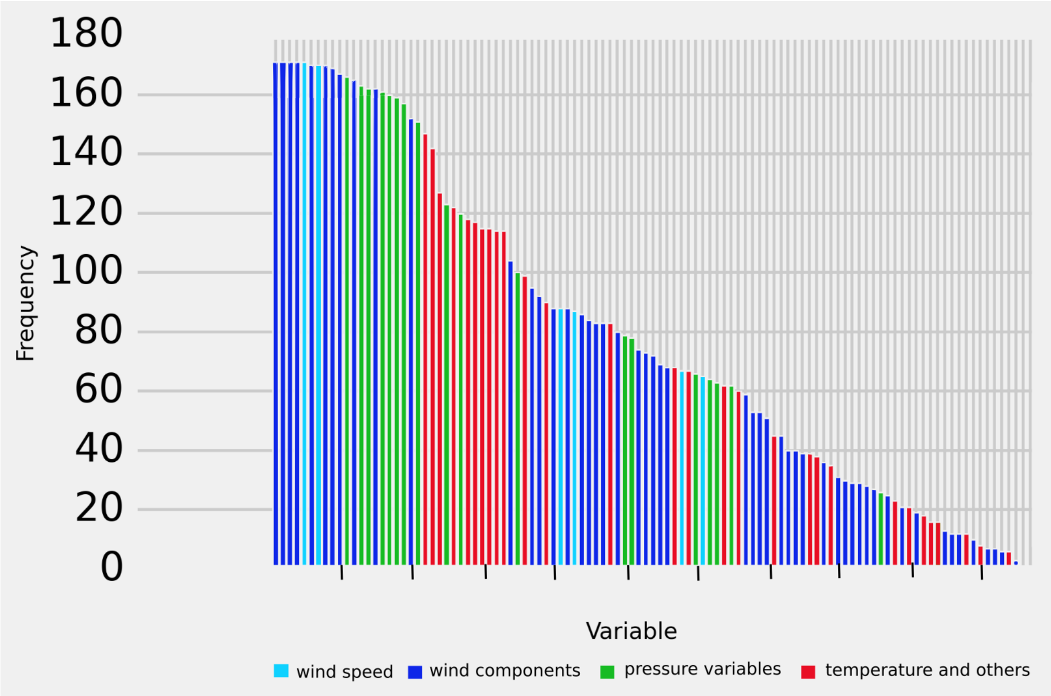



**Fig. 7** Frequency of occurrence of top 50 variables for all the stations in France. Each vertical bar corresponds to one explanatory variable. For readability, we have not included abbreviated names of the variables, but indicate with colors the categories of variables. Note that the explanatory variables based on pressure (to be more precise on the geopotential taken on isobaric surfaces) include the horizontal gradients. These are very close to geostrophic wind, hence to wind.

A model based on a more generic approach named $RF_C$ with 50 explanatory variables common to all stations was carried out and it was found to perform as well as $RF_B$ (Figure 8). To investigate how much the list of variables can be shortened, another $Random\ forest$ model $RF_{C25}$ with 25 most important explanatory variables was set up. At this point, we began to degrade the performance marginally: $RF_{C25}$ is as good as $RF_A$ with just 1% degradation in RMSE overall. However, going down to $RF_{C10}$ with 10 most important variables not only degrades the RMSE by 8% and correlation by 2%, but also increases the IQR of RMSE and correlation by 13% and 11% respectively (refer to Tables 7 and 8). Nonetheless, $RF_{C10}$ performs better than all the parametric models described in Sect. 3.2.

Further analyzing the list of explanatory variables, we found that the wind speed at 100 m (F100), wind speed at 10 m (F10) and bulk wind shear between 10 m and 100 m (DF) are the 3 most significant variables that bring in key information from the synoptic flow at any given location. Accordingly, another model $RF_{C3I}$ with only three variables (F10, F100 and DF) was set up. The results turned out to be more nuanced compared to the parametric models as can be seen from Figure 8. The performance of a $Linear\ regression$ model with the same 3 important explanatory variables is poorer than $RF_{C3I}$ (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 performances, but that a reduction to only 3 explanatory variables is excessive and comes at the cost of degraded performances.

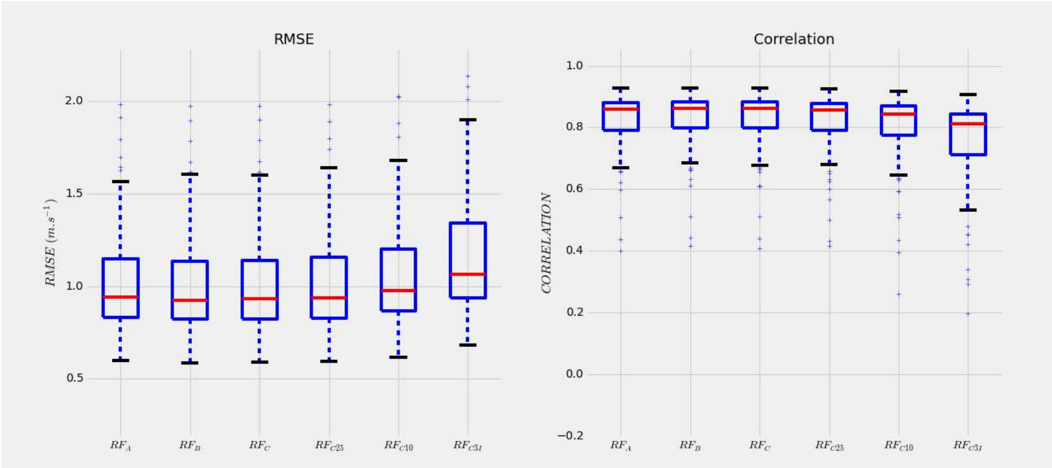



**Fig. 8** Boxplot of RMSE and correlation of various $RF$ models for all the stations in France

**Table 7** Quartiles of the RMSE of various $RF$ models from the boxplot (Figure 8), in $m\,s^{-1}$. The extreme values are extreme values for the whole dataset, ie including outliers.

| Model | Min | Q1 | Median | Q3 | Max | IQR |
|---|---|---|---|---|---|---|
| $RF_A$ | 0.60 | 0.84 | 0.95 | 1.15 | 1.60 | 0.31 |
| $RF_B$ | 0.60 | 0.83 | 0.94 | 1.15 | 1.63 | 0.32 |
| $RF_C$ | 0.60 | 0.83 | 0.94 | 1.14 | 1.62 | 0.31 |
| $RF_{C25}$ | 0.61 | 0.84 | 0.96 | 1.16 | 1.65 | 0.32 |
| $RF_{C10}$ | 0.64 | 0.91 | 1.02 | 1.26 | 1.70 | 0.35 |
| $RF_{C3}$ | 0.72 | 1.00 | 1.12 | 1.38 | 1.92 | 0.38 |

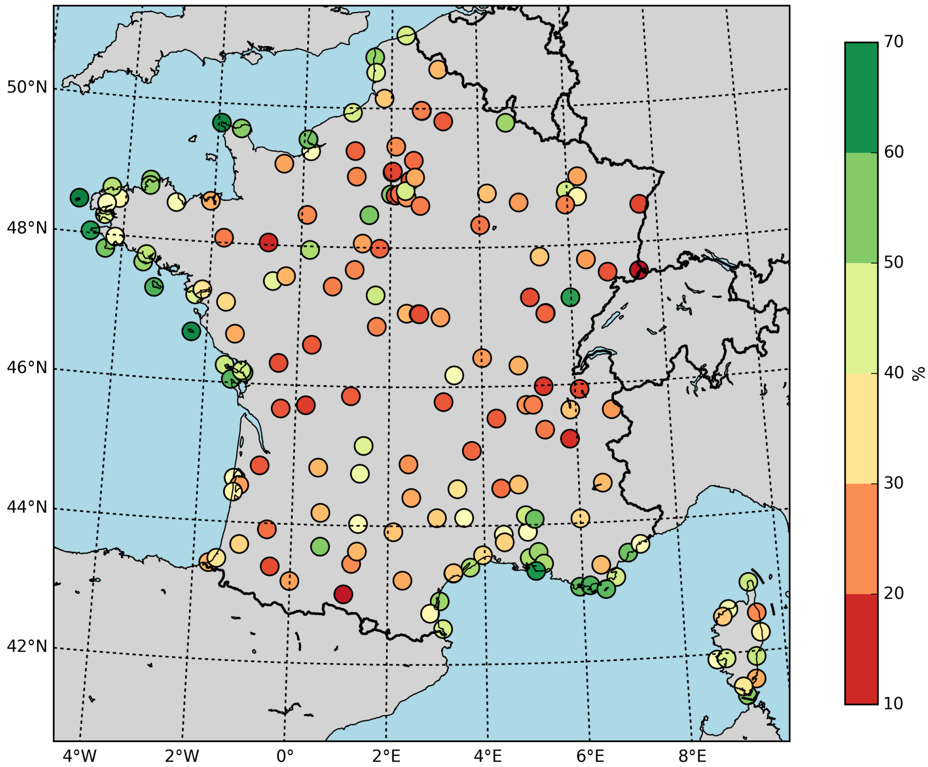
**Table 8** Quartiles of the correlation of various $RF$ models from the boxplot (Figure 8), in $m\,s^{-1}$. The extreme values are extreme values for the whole dataset, ie including outliers.

| Model | Min | Q1 | Median | Q3 | Max | IQR |
|---|---|---|---|---|---|---|
| $RF_A$ | 0.66 | 0.79 | 0.85 | 0.88 | 0.93 | 0.09 |
| $RF_B$ | 0.68 | 0.80 | 0.86 | 0.88 | 0.93 | 0.08 |
| $RF_C$ | 0.67 | 0.80 | 0.86 | 0.88 | 0.92 | 0.08 |
| $RF_{C25}$ | 0.66 | 0.79 | 0.85 | 0.87 | 0.92 | 0.08 |
| $RF_{C10}$ | 0.62 | 0.76 | 0.83 | 0.86 | 0.91 | 0.10 |
| $RF_{C3}$ | 0.51 | 0.69 | 0.79 | 0.82 | 0.88 | 0.13 |

Regarding the spatial distribution, the percentage change in RMSE and correlation of $RF_{C25}$ model with respect to $ECMWF$ is shown in Figure 9. From Figure 9a it can be noticed that the RMSE of inland stations in the north of France improves by 30% on average. The stations in the inland South have a mean improvement in RMSE of 40%. The highest improvements of up to 80% could be recognized on coastal stations in the West, the South and Corsica. From Figure 9b, the correlation follows a similar pattern to RMSE with stations in the inland north and inland south showing an average improvement of 15% and 22% respectively. The coastal stations display an average improvement in correlation of 60%.

In conclusion, $RF_A$ model used an unnecessarily long list of explanatory variables. This was not detrimental to its performance, but needlessly cumbersome. The performance could be slightly improved with the $RF_B$ model with 50 station specific explanatory variables. The model $RF_C$ with 50 common explanatory variables performs as good as $RF_B$ but is generic in nature. $RF_{C25}$ is simple and robust with just 25 important explanatory variables and is comparable to $RF_A$ in performance.

However, with fewer explanatory variables, $RF_{C25}$ is not quite as good as $RF_C$. Hence, $RF_{C25}$ appears as a compromise between performance and simplicity. It is instructive to analyze the list of 25 explanatory variables retained.



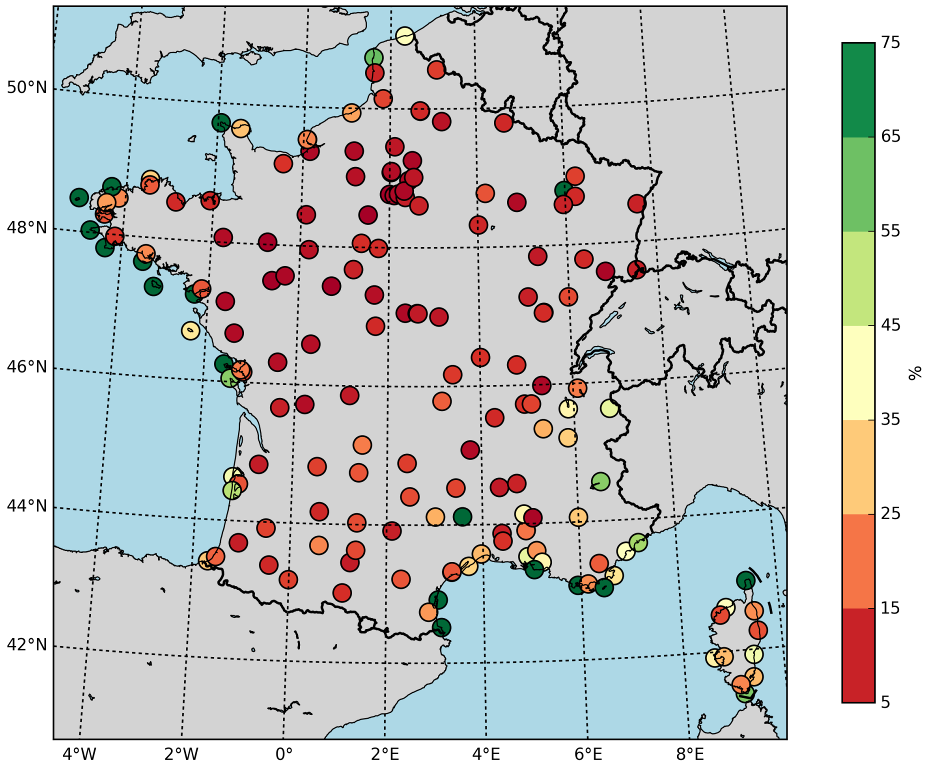




**Fig. 9** Percentage change in RMSE of $RF_{C25}$ model with respect to $ECMWF$ analyses (top), and percentage change in correlation of $RF_{C25}$ model with respect to $ECMWF$ analyses (bottom).

### 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:

- All information (components and speed) on the 10 m and 100 m wind (6 variables),
- the wind shear between 10 $m$ and 100 $m$ (1 variable),
- and the components of 500 $hPa$ wind (2 variables).

Top 25 list:

- Added to the previous list are the wind components at 850 $hPa$ and 925 $hPa$ (4 variables),
- the gradients of geopotential at 925 $hPa$, 850 $hPa$ and 500 $hPa$ (6 variables),
- gradients of mean sea level pressure (2 variables),
- skin temperature,
- temperature at 2 $m$,
- the boundary layer height,
- and one of the gradients of surface pressure.

The subsequent 10 variables include the temperature and boundary layer parameters. In the following appear a few divergence and vorticity variables. Even though the gradients of geopotential that are dominating the second ten list indirectly represent the geostrophic wind components at the respective pressure levels which are in the top ten list, $RF_{C10}$ model did not perform as good as $RF_{C25}$. This suggests that the other variables carry significant information.

To conclude, it is striking that the most relevant variables are almost all wind variables (wind or geostrophic wind). 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 is not the case, which can be explained as follows: the model already describes rather well the wind, and the shear already encompasses the relevant information on the stratification and mixing in the boundary layer, and/or 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 in RMSE and correlation over ECMWF, Figure 9 gave the impression of formulation of three clusters: inland north, inland south, and coastal. We attempt to provide a statistical confirmation in the following section. Another issue that emerged during the study is the influence of time of the day on the errors made by machine learning models. This issue is addressed in Sect. 5.4.

### 5.1 Bias

It was chosen above to quantify the performance of the machine learning models using the RMSE and correlation as complementary tools. Nonetheless, it is important to probe how much of the RMSE results from a reduction of a bias present in the ECMWF output. For this purpose, the bias of the ECMWF surface wind was calculated and is shown in figure 10. It is apparent that the largest values of RMSE (Fig. 2) correspond to the largest values of bias. There is mostly a positive bias over coastal stations, amounting typically to nearly half of the RMSE. There are also a few inland stations displaying a significant negative bias, corresponding to unusually large RMSE for inland stations. Over the whole set of stations, the bias is on average 0.47 $m\,s^{-1}$, amounting to slightly more than a quarter of the RMSE (1.74 $m\,s^{-1}$). For individual stations, the biases range from -1.61 to 2.50 $m\,s^{-1}$.

As expected, the machine learning models prove very efficient at removing the bias. For illustration, the bottom panel of figure 10 displays the bias for $RF_{C25}$, which is uniformly negligible. The average bias is 0.004 $m\,s^{-1}$, with values for individual stations ranging from -0.02 to 0.04 $m\,s^{-1}$.

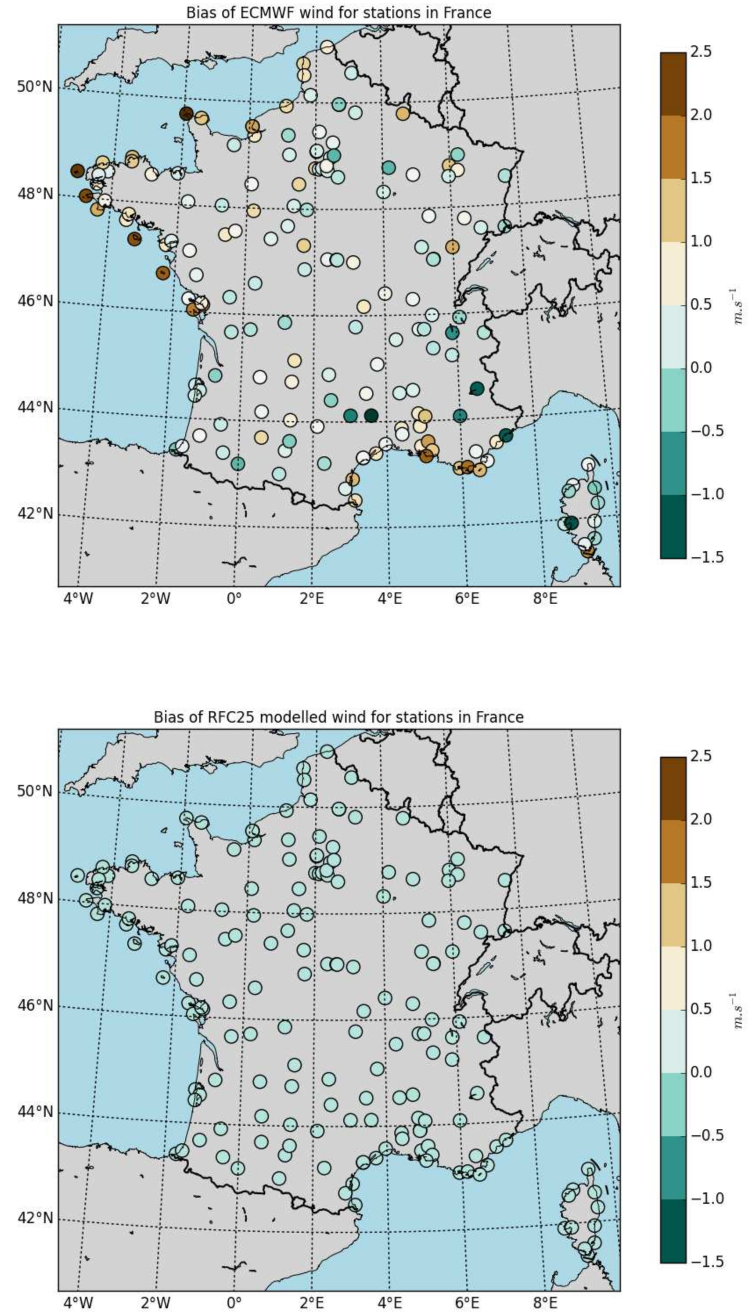



**Fig. 10** Bias in the surface wind for the ECMWF (top panel) and for the estimated surface wind using the $RF_{C25}$ (bottom panel).

## 5.2 Altitude

As a preliminary attempt, we tried to look for a link between the percentage improvement in RMSE and correlation with the altitude of the station. No statistical evidence for such a link was found. As the local topography has a significant effect on the small scale variations of surface winds, we searched for a relation between the small scale gradients of 2 $km$ topography around each station and model performance. We found no clear link between the gradients of altitude and the percentage improvement in RMSE and correlation. We further elaborated our previous approach by taking into account the variance of topography around each station. We achieved this by considering the altitude at 0.2km, 0.5km, 1km, 1.5km, 2km, 3km, and 5km along north, south, east, and west directions around each station and computing the overall variance of altitude. No clear link between the improvement in RMSE and correlation with the altitude parameters was discovered.

### 5.3 Cluster

Independently, unsupervised classification using $k - means\ clustering$ was also performed by feeding RMSE and correlation of $ECMWF$, and percentage change in RMSE and correlation of $RF_{C25}$ over $ECMWF$ as explanatory variables. Nothing conspicuous came out of this approach towards clustering. More work is needed in this direction as finding clusters would help in approximating the error made by $ECMWF$ at other locations with similar topographic variations.

### 5.4 Diurnal

Inspection of the error at specific stations suggested that a diurnal cycle of 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, wind). To illustrate this diurnal cycle, the Probability Density Functions of errors for the four different analysis times (00, 06, 12, and 18 UTC) are shown in Figure 11 at le Havre Octeville station and for the ECMWF output. There clearly are biases that vary with the time of day. The signs of these biases was 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 point this out as a direction for further exploration, and for which a better knowledge of the modeling system and its limitations may be particularly beneficial.

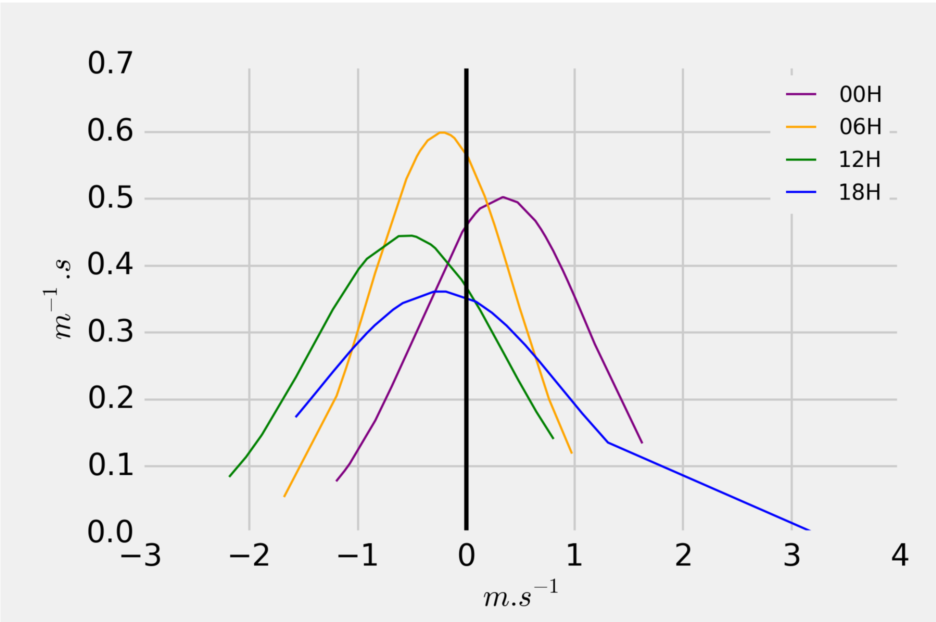



**Fig. 11** Probability distribution of error of the $RF_{C25}$ modeled 10 m wind at various hour indices for the station Le Havre-Octeville

### 5.5 Application to Other Variables

The methodology described in the present paper was applied to surface wind speed because of a strong demand from the wind energy sector for better estimates of surface winds. It is not specific to wind however and could apply to other quantities. For wind itself, it has been applied in preliminary tests to the wind components as an intermediate step before calculating the wind speed. This did not provide a gain for the end caclculation of the wind speed and was not pursued. It could also be used for wind direction, for which statistical estimators can also be used despite its cyclic character (e.g. [Yam82]). Variables other than wind, notably temperature, could be estimated with the above methoology. However, the errors of NWP models on temperature are less of a concern than for winds. The RMSE and correlation of the temperature directly output from the ECMWF was calculated for all the 171 stations (not shown). The average of RMSE is 1.45 K with a standard deviation of 0.84 K, indicating strong variations among stations. Indeed, for individual stations the RMSE ranges from 0.77 to 7.96 K. Excluding four stations which appear as outliers brings the average RMSE down to 1.34 K, with a standard deviation of 0.39 K. The average correlation is 0.98, the weakest correlation being 0.89. 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 10 m wind speed at 171 stations in France. We have shown that post-processing using machine learning models could provide significant improvements over the performance of the NWP model alone. Before reporting our conclusions in Sect. 7, we need to consider an essential question hitherto left aside: in all that precedes, the NWP outputs were extracted from analyses. In practice, it is *forecasts* that will be of use for wind energy operators. Does the relationship identified between model outputs and observed winds 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 intends to improve the forecasts of the surface winds from the outputs of the $ECMWF$ model, using the same post-processing as described in previous sections. Note that this will provide 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 00:00 and 12:00 UTC and each run forecasts the weather up to 10 days. We limit this study to the station Le Havre-Octeville (already used in previous Sect. 3.1). Appropriately, the $ECMWF$ forecast data were retrieved at lead times of 0H, 3H, 6H, 12H, and 24H where 0H corresponds to that of the analyses. The machine learning models used to reconstruct the wind from these forecasts are the same as described and used previously: 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, figure 12 shows the RMSE and correlation of the 10 $m$ winds from $ECMWF$ forecasts at various lead times for the station Le Havre-Octeville in France. As seen previously, the RMSE is rather large (nearly 2.5 $m\,s^{-1}$), and it remains fairly constant over the first 24 hours of the forecast.

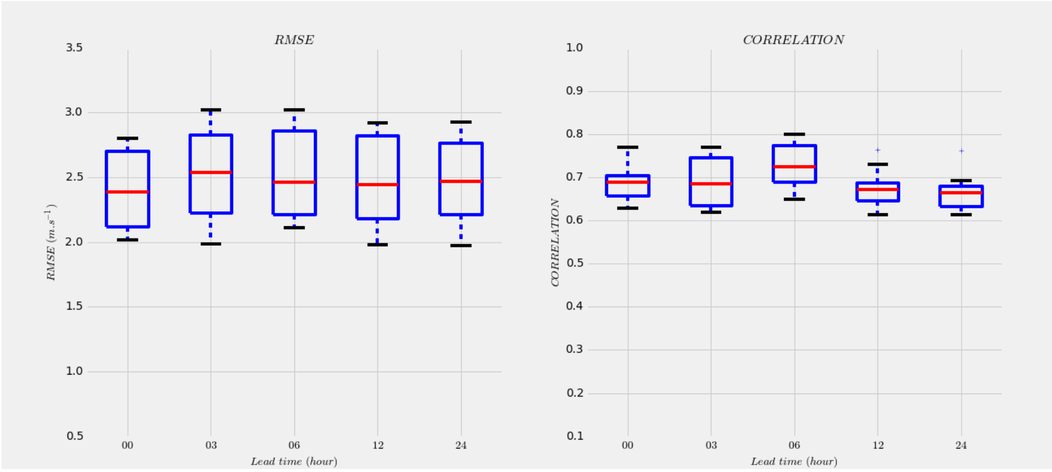



**Fig. 12** RMSE and correlation of the $10m$ $ECMWF$ forecast winds at various time horizon for the station Le Havre-Octeville.

Now, we apply the $RF_{C25}$ model, trained on the analyses as described in Sect. 4, to the outputs of the $ECMWF$ forecasts at lead times from 3 to 24 hours. The RMSE and correlation of the reconstructed wind are shown in Figure 13. Strikingly, the RMSE is dramatically reduced (down to less than 1.2 $m\,s^{-1}$, with a very narrow spread): the average improvement in RMSE and correlation over all the lead times is about 55% and 21% respectively. These improvements are simply consistent with those obtained with Random Forests from the outputs 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 hours; this could be explored if the investigation at other stations confirmed it to be a robust feature, but is beyond the scope of the present study. The message to retain here is that the improvements carry over to forecasts, and that for lead times up to 24 hours 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.

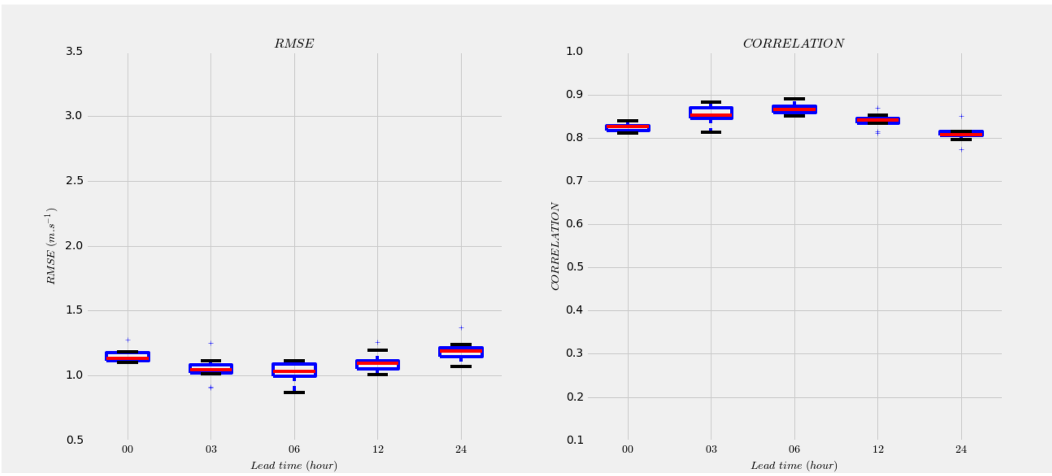



**Fig. 13** RMSE and correlation of $10m$ winds of the $RF_{C25}$ model at various time horizon for the station Le Havre-Octeville

## 7 Conclusion

In this study, we used several parametric and no-parametric machine-learning methods to estimate the surface wind speed from the analyses of the $ECMWF$ model over 171 stations in France. Two issues were particularly emphasized: first, the use and comparison of both parametric methods (multi-linear regression, as in a majority of Model Output Statistics (MOS) practices) and machine learning methods (notably random forests), and second, the identification of model variables that carried most information for the estimation of the surface winds.

The $ECMWF$ model estimates well the 10 $m$ wind speed 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 RMSE and correlation 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 wind, 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 10 $m$ wind, relative to the ECMWF direct model output, as intended. All the parametric models (Linear regression) show a similar performance with an average decrease of 25% for RMSE and increase of 8% for correlation. Tree based non-parametric models (Random forest and Gradient boosting) show the best performance with a mean decrease of 33% for RMSE and increase of 15% for correlation. The KNN model, being not only non-parametric, but also data sensitive, gave intermediate results. The highest improvements in RMSE and correlation 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 technique 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 median of RMSE for models $RF_C$, $RF_{C25}$, $RF_{C10}$ and $RF_C3$ are respectively 0.94, 0.96, 1.02 and 1.12 $m\,s^{-1}$). Hence, the $random\ forest$ model with 25 variables common to all stations ($RF_{C25}$) appeared to be the best compromise between performance and simplicity. A generic list of 25 most significant variables that could be used to predict wind at any location was proposed. It is striking to note that the most relevant variables are almost exclusively wind variables (wind or geostrophic wind). 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 RMSE and correlation, there seems to appear a geographical pattern (with highest 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 made by 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 would call for further, more systematic investigation.

The present study confirms, for the estimation of surface winds, the relevance of machine learning models such as random forests, in agreement with the findings and choices of [ZBMS16]. These authors, in the context of providing improved, gridded data for surface winds, used random forests and explored strategies for obtaining gridded surface winds 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 with forecasts in Sect. 6 opens the way for further studies to apply these models for forecasts, notably for wind energy, using 100 m winds. Another important source of information to tap into are outputs from NWP at higher resolution. The French meteorological agency, Météo-France, produces forecasts for mainland France at a higher spatial resolution ($dx = 1.3$ km presently). Investigating the performance of machine learning models using input from such higher resolution model constitutes a topic for further research.

**Acknowledgements** This research was supported by ANR project FOREWER (ANR-14-CE05-0028).

## References


APM+18. B. Alonzo, R. Plougonven, M. Mougeot, A. Fischer, A. Dupré, and P. Drobinski. *Forecasting and risk management for renewable energy*, chapter From Numerical Weather Prediction outputs to accurate local surface wind speed: statistical modeling and forecasts. Springer, 2018.

BM05. J.A. Baars and C.F. Mass. Performance of National Weather Service forecasts compared to operational, consensus and wieghted model output statistics. *Wea. Forecast.*, 20:1034–1047, 2005.

BTB15. P. Bauer, A. Thorpe, and G. Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525:47–55, 2015.

Cha14. W.-Y. Chang. A literature review of wind forecasting methods. *Journal of Power and Energy Engineering*, 2:161–168, 2014.

dRK04. W.C. de Rooy and K. Kok. A combined physical-statistical approach for the downscaling of wind speed. *Weather and forecasting*, 19:485–495, 2004.

DvLD13. A. Devis, N.P.M. van Lipzig, and M. Demuzere. A new statistical approach to downscale wind speed distributions at a site in northern Europe. *J. Geophys. Res. Atmos.*, 25:2272–2283, 2013.

FLMM12. A.M. Foley, P.G. Leahy, A. Marvuglia, and E.J. McKeogh. Current methods and advances in forecasting of wind power generation. *Renewable Energy*, 37:1–8, 2012.

GL72. H.R. Glahn and D.A. Lowry. The use of model output statistics (MOS) in objective weather forecasting. *J. App. Meteor.*, 11:1203–1211, 1972.

GWHT13. James Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer, 2013.

HJ+18. T. Haiden, M. Janousek, , J.-R. Bidlot, R. Buizza, L. Ferranti, F. Prates, and F. Vitart. Evaluation of ecmwf forecasts, including the 2018 upgrade. *ECMWF Technical Memo.*, 831, October 2018.

HOP12. V. Horlacher, S. Osborne, and J.D. Price. Comparison of two closely located meteorological measurement sites and consequences for their areal representativity. *Boundary Layer Meteorology*, 142:469–493, 2012.

Kal03. Eugenia Kalnay. *Atmospheric modeling, data assimilation and predictability*. Cambridge University Press, 2003.

KSHK11. J.-H. Kang, M.-S. Suh, K.-O. Hong, and C. Kim. Development of updateable model output statistics (UMOS) system for air temperature over South Korea. *Asia-Pac. J. Atmos. Sci.*, 47:199–211, 2011.

LPZI14. L. Lazic, G. Pejanovic, M. Zivkovic, and L. Ilic. Improved wind forecasts for wind power generation using the Eta model and MOS (Model Output Statistics). *Energy*, 73:567–574, 2014.

MGW18. J. Mejia, M. Giordano, and E. Wilcox. Conditional summertime day-ahead solar irradiance forecast. *Solar Energy*, 163:610–622, 2018.

MM16. John Paul Muller and Luca Massaron. *Machine learning for dummies*. John Wiley & Sons, 2016.

RGC13. M. Ranaboldo, G. Giebel, and B. Codina. Implementation of a model output statistics based on a meteorological variable screening for short-term wind power forecasts. *Wind Energy*, 16:811–826, 2013.

Rid13. Bob Riddaway. Newsletter no. 136 - summer 2013. 07 2013.

SDVN09. T. Salameh, P. Drobinski, M. Vrac, and P. Naveau. Statistical downscaling of near-surface wind over complex terrain in southern France. *Meteorol. Atmos. Phys.*, 103:253–265, 2009.

SKV05. M.J. Schmeits, K.J. Kok, and D.H. Vodelezang. Probabilistic forecasting of (severe) thunderstorms in the Netherlands using model output statistics. *Wea. Forecast.*, 20:134–148, 2005.

SLV11. A. Smith, N. Lott, and R. Vose. The Integrated Surface Database: Recent Developments and Partnerships. *Bull. Am. Meteor. Soc.*, 92:704–708, 2011.

STG12. N. Schuhen, T.L. Thorarinsdottir, and T. Gneiting. Ensemble model sutput statistics for wind vectors. *Monthly Weather Review*, 140:3204–3219, 2012.

Tib96. R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Stat. Soc. Series B*, 58:267–288, 1996.

TU14. A. Tascikaraoglu and M. Uzunoglu. A review of combined approaches for prediction of short-term wind speed and power. *Ren. Sust. Energy Rev.*, 34:243–254, 2014.

WD13. R.L. Wilby and C.W. Dawson. The Statistical Downscaling Model: insights from one decade of application. *Int. J. Climatol.*, 33:1707–1719, 2013.

WGH11. X. Wang, P. Guo, and X. Huang. A review of wind power forecasting models. *Energy Procedia*, 12:770–778, 2011.

WV02. L.J. Wilson and M. Vallée. The Canadian Updateable Model Output Statistics (UMOS) systemm: Design and Development tests. *Wea. Forecast.*, 17:206–222, 2002.

Yam82. R.J. Yamartino. A comparison of several 'single-pass' estimators of the standard deviation of wind direction. *J. Clim. App. Met.*, 23:1362–1366, 1982.

ZBMS16. M. Zamo, L. Bel, O. Mestre, and J. Stein. Improved gridded wind speed forecasts by statistical postprocessing of numerical models with block regression. *Weather and Forecasting*, 31:1929–1945, 2016.

ZMAP14. M. Zamo, O. Mestre, P. Arbogast, and O. Pannecouke. A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production, part I: Deterministic forecast of hourly production. *Solar Energy*, 105:792–803, 2014.