From Numerical Weather Prediction outputs to accurate local surface wind speed : statistical modeling and forecasts.

Alonzo B.[1,2], Plougonven R.[1], Mougeot M.[2], Fischer A.[2], Dupré A.[1] and Drobinski P.[1]

[1]LMD/IPSL, cole Polytechnique, Université Paris Saclay, ENS, PSL Research University, Sorbonne Universités, UPMC Univ Paris 06, CNRS, Palaiseau France
[2]Laboratoire de Probabilités et Modèles Aléatoires, Université Paris Diderot - Paris 7, Paris, France.

November 8, 2017

Abstract

Downscaling a meteorological quantity at a specific location from outputs of Numerical Weather Prediction models is a vast field of research with continuous improvement. The need to provide accurate forecasts of the surface wind speed at specific locations of wind farms has become critical for wind energy application. While classical statistical methods like multiple linear regression have been often used in order to reconstruct wind speed from Numerical Weather Prediction model outputs, machine learning methods, like Random Forests, are not as widespread in this field of research. In this paper, we compare the performances of two downscaling statistical methods for reconstructing and forecasting wind speed at a specific location from the European Center of Medium-range Weather Forecasts (ECMWF) model outputs. The assessment of ECMWF shows for 10m wind speed displays a systematic bias, while at 100m, the wind speed is better represented. Our study shows that both classical and machine learning methods lead to comparable results. However, the time needed to pre-process and to calibrate the models is very different in both cases. The multiple linear model associated with a wise pre-processing and variable selection shows performances that are slightly better, compared to Random Forest models. Finally, we highlight the added value of using past observed local information for forecasting the wind speed on the short term.

Keywords

Local wind speed, Downscaling, Statistical modeling, Numerical Weather Prediction model, Wind speed forecasts.

1 Introduction

The wind energy sector has seen a very sharp growth in recent years. According to the Global Wind Energy Council (GWEC), 54GW has been installed in 2016, corresponding to an increase of 12.6% of the total installed capacity ([GWE, 2016]). Worldwide, the number of wind farms increases each year and feeds the electrical network with a larger amount of energy. For instance, in 2016, France has seen its highest capacity growth rate ever recorded. This sharp increase of connected wind power has for instance allowed the network to receive 8.6 GW from wind power plants, on November 20th, corresponding to 17.9% of the energy produced this day ([RTE, 2016]). The need to have access to accurate wind forecasts on several timescales is thus becoming crucial for the wind energy producer and grid operator, in order to anticipate the energy production, to plan maintenance operations and to handle balance between energy production and consumption. Changing regulations of the energy market with the end of feeding-in tariffs make this anticipation vital for wind energy producers. Finally, a related but different topic consists in the estimation of the wind resource of its long-term (multi-year) variability and trends mainly for prospecting purposes.

The increasing need for accurate forecasts of the surface wind speed fortunately comes with the improvement of the Numerical Weather Prediction models (NWP) describing and forecasting atmospheric motions. Indeed, they constitute a key source of information for surface wind speed forecasts all the more so as their realism, accuracy and resolution have increased steadily over the years [Bauer et al., 2015].

Nevertheless, these models are not necessarily performing uniformly well for all atmospheric variables. Their astonishing performances are evaluated on variables such as mid-tropospheric pressure which reflect the large-scale mass distribution, which is effectively well understood physically (e.g. [Vallis, 2006]) and efficiently modeled numerically. Variables tied to phenomena occurring on smaller scales (such as cloud-cover or near-surface winds) depend much more directly on processes that are *parameterized* (e.g, not resolved). In contrast to large-scale motions (governed by the Navier-Stokes equations), parameterizations are generally partly rooted in physical arguments, but also in large part empirical. When comparing output from a numerical model to a local measurement, there will therefore always be several sources of error: representativity error (contrast between the value over a grid-box and the value at a specific point), numerical error (even if we were describing only processes governed by well-established physical laws, discretization is unavoidable), and error tied to the physics described (because processes, especially parameterized ones, are not well modeled). To reduce representativity error and to better represent small-scale processes, in particular those tied to topography and surface roughness, one strategy consists in downscaling with models that describe the atmospheric flow on finer scales (e.g [Wagenbrenner et al., 2016]). One disadvantage of this approach is the numerical cost, and one limitation is the need for finer observations to initialize the state of the atmosphere, if details of the flow other than those directly implied by the topography and surface condition are sought for.

Strategies to estimate surface winds, or other meteorological variables, from the output of Numerical Weather Prediction models (NWP) or climate models have been developed in several contexts, with different motivations, and leading to different methodologies and applications.

Model Output Statistics (MOS) has been developed in weather forecasting for several decades to estimate the *weather related* variability of a physical quantity, based on NWP model output [Glahn and Lowry, 1972]. NWP models perform now very well in predicting large-scale systems. Relations thus can be derived to link the latters to local variables at an observation site. Linear models are generally used, with the outcome now expanded over a wider area than only the location of stations where observations are available [Zamo et al., 2016].

In the context of climate change, downscaling a meteorological quantity at a given location in order to produce time series which have plausible statistical characteristics under climate change has for long been investigated [Wilby et al., 1998]. The challenge is here to capture appropriately the relation between large-scale flow (as it can be described by a model with a moderate or low resolution) and a variable at a specific location (e.g. wind, temperature, precipitation) and then use climate models to provide a description of the large-scale atmospheric state under climate change. Local time series with appropriate variability and consistent with this large-scale state of the atmosphere are then generated, e.g. [Salameh et al., 2009, Maraun et al., 2010, Wilby and Dawson, 2013].

Wind energy domain is nowadays a very active branch in downscaling techniques because of the need for accurate forecasts at specific location of a wind farm. For describing winds close to the surface, 10m wind speed is often a convenient variable as it has been for decades a reference observed variable and also now a reference NWP model output. In the case of wind energy, the wind speed then needs to be extrapolated at the hub height to have access to wind power, leading to an increase of the error on the predicted power ([Kubik et al., 2011], [Howard and Clark, 2007], [Mohandes et al., 2011]). Wind speed at the hub height (typically 100m) is a variable of interest as it allows to avoid vertical extrapolation errors ([Cassola and Burlando, 2012]), but it is rarely available in observations. Different outputs of NWP models can be used as explanatory variables of the near surface wind speed. It seems that there is no strong consensus on the predictors to use, mainly because relations between predictors and predictand should differ from one location to the other. However, different studies have shown the importance of a certain set of variables to predict surface wind speed. Amongst them, markers of large-scale systems (geopotential height, pressure fields) and boundary layer stability drivers (surface temperature, boundary layer height, wind and temperature gradient) can be cited ([Salameh et al., 2009], [Devis et al., 2013], [Davy et al., 2010]). In terms of methodology, several models have already been studied, including Linear regression, Support Vector Models (SVM) or Artificial Neural Network (ANN) ([Jung and Broadwater, 2014], [Soman et al., 2010]).

The model of the European Center for Medium-range Weather Forecasts (ECMWF) has reached a resolution of about 9km in the horizontal. In addition, ECMWF analyses and forecasts now give access to 100m wind speed output, developed mainly for wind energy applications. If we can be very confident in the ability of NWP models to represent several variables, some others may not be so reliable. This is especially the case for surface variables such as 10m and 100m wind speed. Consequently, using the robust information given by some variables to correct surface wind speed is straightforward. We have access to surface wind speed observed at 10m, 100m over a long period of 5 years at SIRTA observation platform [Haeffelin et al., 2005]. The aim of this project is, in particular, to explore how different statistical models perform in forecasting the 10m and 100m wind speed using informations of ECMWF analyses and forecasts outputs at different horizons. We choose multiple linear regression because it is a widely used technique, and Random Forests which have not been, to our knowledge, deeply studied in the framework of downscaling surface wind speed. For multiple linear regression, variable selection is a very important step for calibrating the statistical models, whereas Random Forests handle variables automatically. Moreover, Random Forests can handle nonlinear relations very well. Therefore, the comparison of those very different statistical models, as well as the information used by each of them, should be very instructive.

The paper is organized in 5 parts. The next section describes together the data and the statistical models to be used. In section 3, the training dataset is explored, and used to calibrate the statistical models. In section 4, forecasts of 10m and 100m wind speed are run to downscale wind speed at the observation site. In the last section, we discuss the results, conclude and give perspectives to this work.

2 Data and Methodology

2.1 Data

Observed Wind speed

In this paper, we use observations of the wind speed at the SIRTA observation platform ([Haeffelin et al., 2005]). Surface wind speed at 10m height from anemometer recording is available at the 5-minutes frequency. The wind speed at 100m height from Lidar recording is available at 10-minutes frequency. Both data span for 5 years from 2011 to 2015. We filter observations by a sinusoidal function over a 6-hour window centered at 00h, 06h, 12h and 18h to obtain a 6-hourly observed wind speed to be compared to the NWP model outputs available at this time frequency. We found that the resulting time series are not sensitive to the filter function. We also try different filtering windows, concluding that 6-hours is the best to compare to the NWP model outputs. Due to some missing data, two final time series of 5049 filtered observations are computed (over 7304 if all data were available).

SIRTA observatory is based 20km in the South of Paris on the Saclay plateau ($48.7°N$ and $2.2°E$). Figure 1 shows the SIRTA observation platform location, marked by the red point on the map, and its close environment. Regarding the relief near SIRTA, observe that a forest is located at about 50m north to the measurement devices. South, buildings can be found at about 300m from the SIRTA observatory. In the East-West axis, no close obstacle are encountered. Further south, the edge of the Saclay plateau

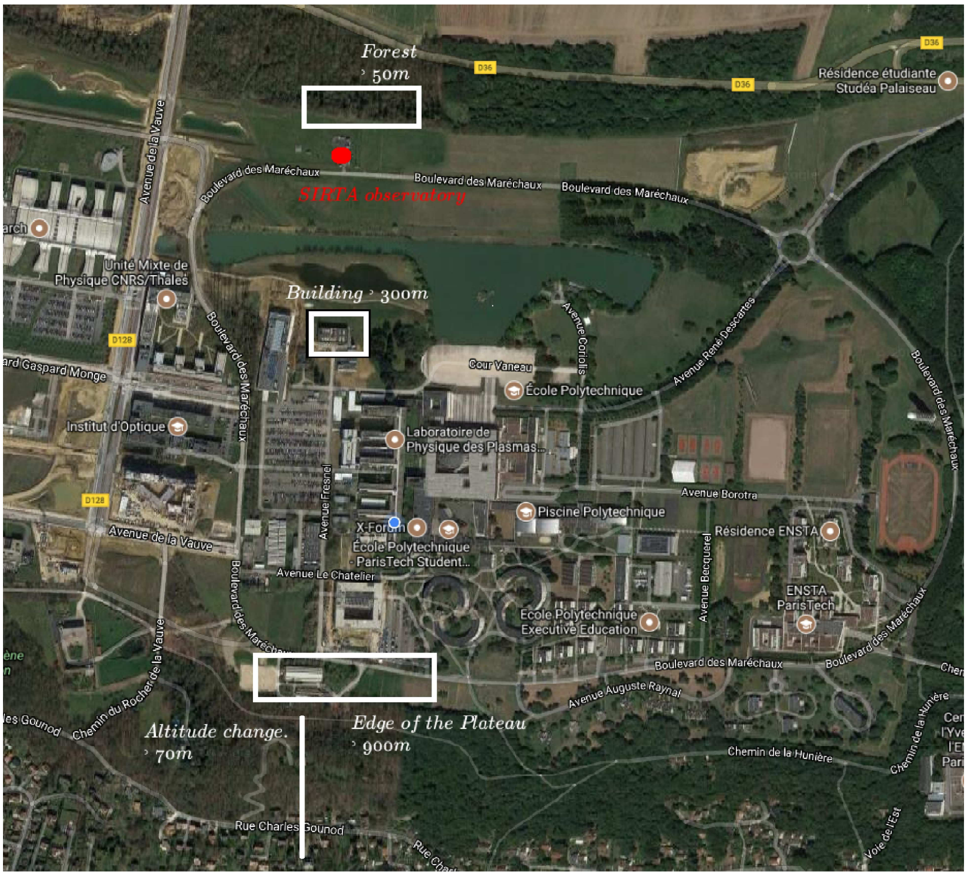

Figure 1: Map of the SIRTA observation platform and its surroundings.

shows a vertical drop of about 70m, from 160m on top to 90m at the bottom.

NWP model outputs - ECMWF Analyses

Variables are retrieved from ECMWF analyses at 4 points around the SIRTA platform. The spatial resolution of ECMWF analyses is of about 16km (0.125° in latitude and longitude). Topography is thus smoothed compared to the real one. As the surface wind speed is very influenced by the terrain, the modeled surface wind speed is not necessarily close to the observed wind speed. The data spans from the 01/01/2011 to 31/12/2015 at the 6-hourly frequency. It is sampled at each date where a filtered sampled observation is available.

The near surface wind speed at a given location can be linked to different phenomena. The large-scale circulation brings the flow to the given location explaining the slowly varying wind speed. The wind speed in altitude, the geopotential height, the vorticity, the flow divergence, sometimes the temperature can be markers of large systems like depressions, fronts, storms, or high pressure systems explaining a large part of the low frequency variations of the surface wind speed (Table 2). At a finer scale, what is happening in the boundary layer is very important to explain the intra-day variations of the wind speed. The state and stability of the boundary layer can be derived from surface variables describing the exchanges inside the layer. Exchanges are driven mostly by temperature gradient and wind shear that develop turbulent flow (Table 3). Thermodynamical variables like surface, skin, and dew point temperatures and surface heat fluxes can also inform on the stability of the boundary layer, as well as its height and dissipation on its state (Table 1). In the end, 20 output variables are retrieved from ECMWF analyses at the 4 points around the SIRTA observatory and at different pressure levels. Note that we restrict the study to local variables (at the location of measurements or in the column above). It might also be possible to take advantage from larger scale information ([Zamo et al., 2016], [Davy et al., 2010]).

Altitude (m)	Variable	Unit	Name
10m/100m	Norm of the Wind speed	$m.s^{-1}$	F
10m/100m	Zonal Wind speed	$m.s^{-1}$	U
10m/100m	Meridional Wind speed	$m.s^{-1}$	V
2m	Temperature	K	T
2m	Dew point Temperature	K	Dp
Surface	Skin temperature	K	skt
Surface	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 1: Surface Variables

Pressure level (hPa)	Variable	Unit	Name
1000hPa/925hPa/850hPa/700hPa/500hPa	Zonal Wind speed	$m.s^{-1}$	U
1000hPa/925hPa/850hPa/700hPa/500hPa	Meridional Wind speed	$m.s^{-1}$	V
1000hPa/925hPa/850hPa/700hPa/500hPa	Geopotential height	$m^2.s^{-2}$	Z
1000hPa/925hPa/850hPa/700hPa/500hPa	Divergence	s^{-1}	Di
1000hPa/925hPa/850hPa/700hPa/500hPa	Vorticity	s^{-1}	Vo
1000hPa/925hPa/850hPa/700hPa/500hPa	Temperature	K	T

Table 2: Altitude Variables

Pressure level (hPa)	Variable	Unit	Name
10m to 925hPa	Wind shear	$m.s^{-1}$	ΔF
10m to 925hPa	Temperature gradient	K	ΔT

Table 3: Computed Variables

ECMWF deterministic forecasts

The year 2015 of deterministic forecasts is retrieved from ECMWF model. A forecast is launched every day at 00:00:00 UTC. The time resolution retained is of 3 hours and the maximum lead-time is 5 days. The same variables as for the analyses are retrieved at the same points around the SIRTA platform.

2.2 Methodology

Our aim is to model the real observed wind speed from outputs of NWP model described above. More specifically, we use ECMWF analyses i.e the best estimate of the atmospheric state at a given time using a model and observations ([Kalnay, 2003]). In what follows, the observed wind speed is the target and the analysed variables are potential explanatory features. Because of the complexity of meteorological phenomena, statistical modeling provides an appropriate framework for corrections of representativity errors and the modeling of site dependent variability. In this context, two main directions may be as usual investigated, *parametric* and *nonparametric* models.

Parametric models assume that the underlying relation between the target variable and the explanatory variables has, relatively to a certain noise, a particular analytical shape depending on some parameters, which need to be estimated through the data. Among this family of models, the linear model with a Gaussian noise is widely used, mostly thanks to its simplicity [Friedman et al., 2001]. Associated to an adequate variable selection, it may be very effective.

Nonparametric models do not suppose in advance a specific relation between the variables: instead, they try to learn this complex link directly from the data itself. As such, they are very flexible, but their performance usually strongly depends on regularization parameters. The family of nonparametric models is quite large: among others, one may cite the nearest neighbors rule, the kernel rule, neural networks, support vector machines, regression trees, random forests... Regression trees, which have the advantage of being easily interpretable, show to be particularly effective when associated to a procedure allowing to reduce their variance as for the Random Forest Algorithm.

Let us describe the linear model and random forests in our context with more details. The linear model supposes a relation between the target Y_t, observed wind speed at time t, and explanatory variables $X_t^1, \dots, X_t^d$, available from the ECMWF, at this time t. For lightening the notation, we omit the index t in the next equation. The linear model is given by

$$Y = \beta_0 + \sum_{j=1}^{d} \beta_j X_j + \varepsilon,$$

where the β_j's are coefficients to be estimated using least-square criterion minimization method, and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ represents the noise. Among the meteorological variables $X_1, \dots, X_d$, some of them provide more important information linked to the target than others, and some of them may be correlated. In this case, the stepwise variable selection method is useful to keep only the most important uncorrelated variables [Friedman et al., 2001]. Denoting by $\hat{\beta}_0, \dots, \hat{\beta}_d$ the final coefficients obtained this way (some of them are zero), the estimated wind $\hat{Y}$ is then given by

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{d} \hat{\beta}_j X_j. \quad (1)$$

An alternative approach to perform variable selection and regularization is to use the Lasso method (see for instance [Tibshirani, 1994]), relying on minimization of the least square criterion penalized by the ℓ^1 norm of the coefficients $\beta_1, \dots, \beta_d$. More specifically, for this model, the predicted wind speed at time t is a linear combination of all the previous variables as in equation (1), the coefficients $\beta_1, \dots, \hat{\beta}_d$ being estimated using the least square procedure, under the constraint $\sum_{j=1}^{d} |\beta_j| \leq \kappa$ for some constant $\kappa > 0$.

Regression trees are binary trees built by choosing at each step the cut minimizing the intra-node variance, over all explanatory variables $X_1, \dots, X_d$ and all possible thresholds (denoted by S_j hereafter). More specifically, the intra-node variance, usually called deviance, is defined by

$$D(X_j, S_j) = \sum_{X_j < S_j} (Y_s - \overline{Y}^-)^2 + \sum_{X_j \geq S_j} (Y_s - \overline{Y}^+)^2,$$

where $\overline{Y}^-$ (respectively $\overline{Y}^+$) denotes the average of the observed wind speed in the area $\{X_j < S_j\}$ (respectively $\{X_j \geq S_j\}$). Then, the selected j_0 variable and associated threshold is given by $(X_{j_0}, S_{j_0}) = \arg\min_{j,S_j} D(X_j, S_j)$. The prediction is provided by the value associated to the leaf in which the observation falls.

To reduce variance and avoid over-fitting, it may be interesting to generate several bootstrap samples, fitting then a tree on every sample and averaging the predictions, which leads to the so-called Bagging procedure [Breiman, 1996]. More precisely, for B bootstrap samples, the predicted power is given by

$$\hat{Y} = \sum_{b=1}^{B} \hat{Y}^b, \tag{2}$$

where $\hat{Y}^b$ denotes the wind speed predicted by the regression tree associated with the b-th bootstrap sample. To produce more diversity in the trees to be averaged, an additional random step may be introduced in the previous procedure, leading to Random Forests, where the best cut is chosen among a smaller subset of randomly chosen variables. The predicted value is the mean of the predictions of the trees, as in equation (2).

3 The relationship between analysed and observed winds

3.1 10m/100m wind speed variability comparison

In this section we compare the observed wind speed at 10m and 100m with the 10m and 100m wind speed output of the ECMWF analyses, respectively. Figure 2 shows the Probability Density Function (PDF) of the wind speed coming from ECMWF analyses and observations, and also for illustration an example of a time series of corresponding wind speeds. It appears that the 10m wind speed from ECMWF analyses displays a systematic bias by overestimating the 10m observed wind speed (Figure 2, a and b). The wind at 100m is comparatively well modeled in terms of variations in the time series, but also in terms of distribution (Figure 2, c and d). It seems that the errors mainly come from the overestimation of peaked wind speeds and the underestimation of low wind speeds (Figure 2, c and d). As 10m wind speed is very influenced by even low topography and surrounding obstacles, which are smoothed or not represented in ECMWF analyses, some of its variations are not well described, and even a quite systematic bias is displayed. The effect of the topography and terrain specificity have less impact on the 100m wind speed, so that it is much better represented in ECMWF analyses.

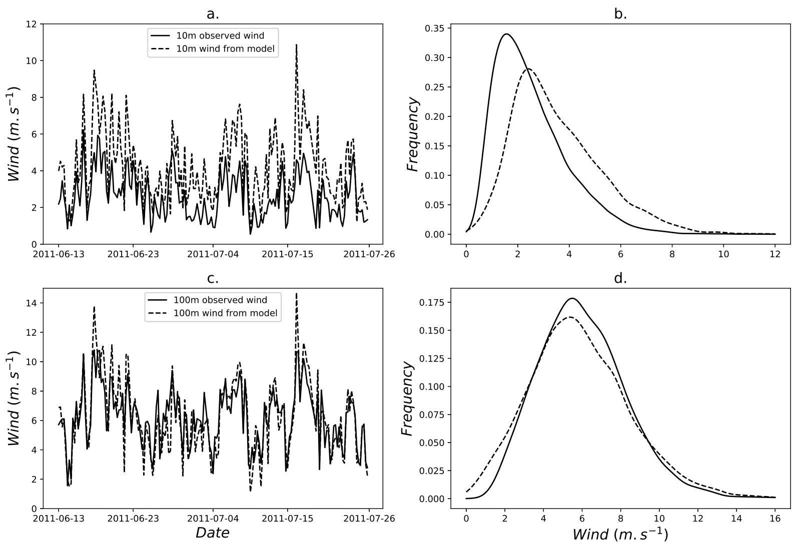

Figure 2: 10m (top) and 100m (bottom) wind speed time series in summer 2011 (panels a and c, respectively) and the respective probability density function estimated over the 5 years sample wind speed (panels b and d).

Periods	Deviation		RMSE		Correlation	
	F10	F100	F10	F100	F10	F100
2011-2015	-1.00	0.14	1.41	1.01	0.82	0.93
2011	-1.19	0.04	1.59	1.06	0.80	0.91
2012	-0.94	0.23	1.31	1.03	0.85	0.92
2013	-1.13	0.06	1.52	0.93	0.82	0.94
2014	-0.88	0.26	1.30	1.00	0.80	0.93
2015	-0.87	0.14	1.30	0.97	0.82	0.94
Winter	-0.97	0.04	1.41	0.97	0.83	0.94
Spring	-1.11	0.27	1.56	1.02	0.71	0.90
Summer	-0.92	0.33	1.31	1.04	0.80	0.91
Fall	-1.04	-0.10	1.36	1.00	0.87	0.93

Table 4: Deviation, RMSE, and Correlation performed by ECMWF analyses for modeling the 10m and 100m wind speed.

The ability of the model to represent the observed wind speed is quantified in Table 4 by the deviation, the Root Mean Square Error (RMSE), and the Pearson correlation which formula are given by equations (3),(4), and (5) respectively.

$$\text{Deviation for the } i^{th} \text{ observation} = (y_i - x_i) \tag{3}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n}} \tag{4}$$

$$\text{Correlation} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{5}$$

where x_i is the wind speed from the NWP model and y_i the observed wind speed ; n is the number of samples (x_i, y_i) and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean) and analogously for $\bar{y}$.

No clear improvement of the ECMWF analyses over the years from 2011 to 2015 can be detected in Table 4. The correlation stays quite constant over the years for both 10m and 100m wind speeds. The Deviation and RMSE seem to decrease for the 10m wind speed but it cannot be confirmed because of the good score performed in 2012. The variations of performance may only come from changes in the predictability of the weather over Europe [Folland et al., 2012]. Seasonal variations of the performance of ECMWF analyses can be seen, especially on the correlation between the observed and modeled wind speed. At both 10m and 100m, the analysed wind speed is better correlated with the observations in winter and fall than in spring and summer. In all cases, the scores shown are better for the 100m wind speed than for the 10m wind speed.

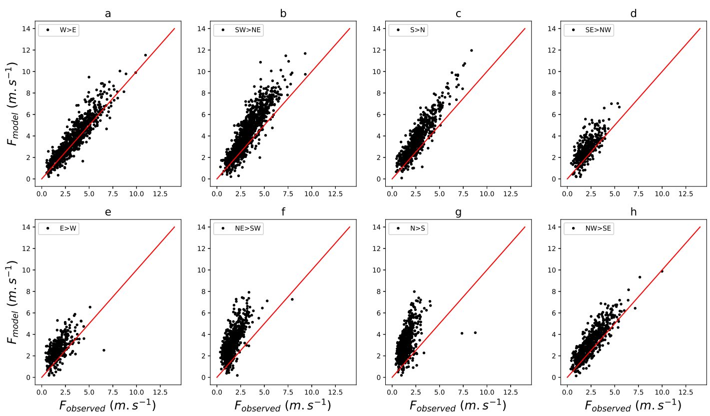

Figure 3: 10m wind speed from ECMWF analyses function of the 10m observed wind speed given cardinal directions. Panels correspond to a direction modeled by ECMWF analyses ; the wind blows from a. West, b. Southwest, c. South, d. Southeast, e. East, f. Northeast, g. North, h. Northwest.

Variations of the performance of the ECMWF analyses in representing the observed wind speed are evidenced by Figure 3. The figure shows the

10m wind speed from ECMWF analyses as a function of the 10m observed wind speed for different directions of the analysed wind speed. It is obvious that the errors made by the numerical model differ regarding the direction of the wind. For instance, when the wind comes from the West (figure 3, a), the correlation is very close to one, but for a wind coming from the North/Northeast (Figure 3, f and e), it is very low, and the model highly overestimates the 10m wind speed. It can be easily linked to the specificity of the terrain. Indeed, when a northerly wind is recorded, it has been blocked by the forest north of the anemometer. The same happened for southerlies with the building situated further and which influence is thus not as substantial as the forest. Figure 4 displays the same as Figure 3 but for the 100m wind speed. It seems that there is no more dependence of the performance of the ECMWF analyses regarding the direction of the 100m wind speed ; it appears to be not significantly impacted by the surrounding forests and buildings.

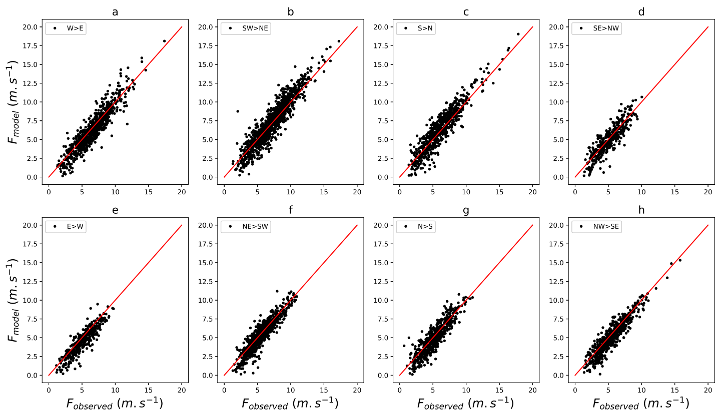

Figure 4: Same as Figure 3 but for 100m wind speed.

3.2 Reconstruction of the 10m/100m observed wind speed using NWP outputs

In the sequel, a k-fold cross validation is performed over 10 different periods taken within the 5-years of analyses and observation. Each time, statistical downscaling models are trained on a given period and applied over the remaining period to reconstruct the 10m and 100m wind speed, so that it results in 10 reconstructions that span the 5 years of data. Table 5 enumerates the statistical downscaling models assessed in this study. Models differ by their types (Linear Regression and Random Forests), the explanatory variable selection, and whether a model is conditionally fitted regarding the direction of the wind speed or not. We evaluate the different statistical models in terms of RMSE and Correlation with the observed wind speed on the reconstruction period.

10m wind speed reconstruction

Model type	Explanatory variables	Direction dependance	Name
Linear	F10	No	LR_F
Linear	All	No	LR_A
Linear	Stepwise	No	LR_{SW}
Linear	Lasso	No	LR_{La}
Linear	F10	Yes	LR_F^{dir}
Linear	All	Yes	LR_A^{dir}
Linear	Stepwise	Yes	LR_{SW}^{dir}
Random Forest	All	No	RF_A
Random Forest	All	Yes	RF_A^{dir}

Table 5: Statistical models used to downscale 10m and 100m wind speed.

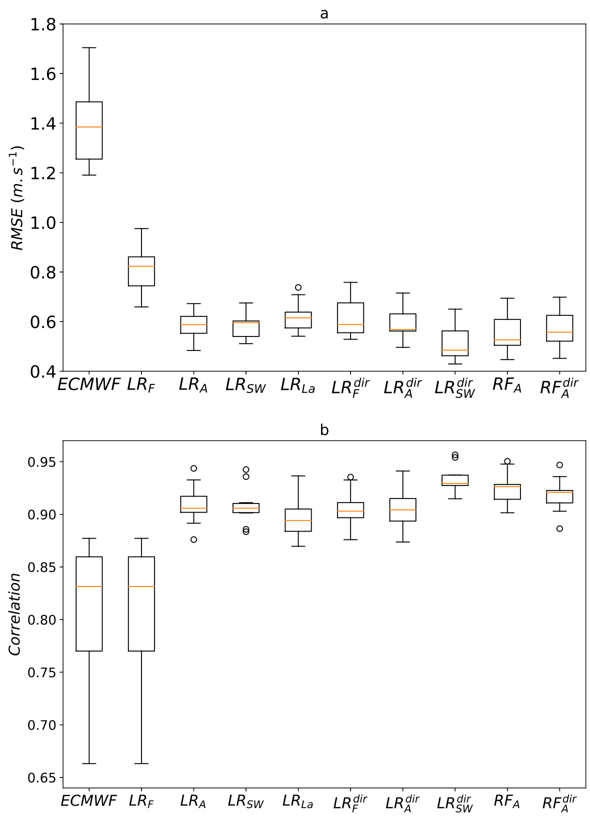

Figure 5: RMSE and Correlation results when reconstructing 10m wind speed with models described in Table 5. The first boxes stand for the ECMWF analyses 10m wind speed.

Figure 5 shows results for the reconstruction of the 10m wind speed. Each box contains the 10th reconstructed k-fold periods. First, by using only wind speed with a linear model LR_F, RMSE is reduced by about 40%, but the correlation stays constant. Adding other variables to linear model (i.e. LR_A, LR_{SW} and LR_{La}) allows to reduce the RMSE by 60%, and to significantly improve correlation from 0.80 to 0.91 between reconstructed wind speed and observed one. Using stepwise selection of variables, the Lasso penalization or all variables does not change results in this case, showing that only a part of the information is useful. Using variable selection as stepwise or ℓ_1 penalization (Lasso) avoids over-fitting. Random Forests models perform slightly better than linear models without defining one given model per cardinal wind directions. Variables selected stepwise are very diverse (wind speed, large scale variables, boundary layer state drivers), while Lasso technique mainly selects wind speed and wind component, thus using redundant information. Analyzing the main variables used by Random Forests shows that this methods seems to put much weight on wind component first, highlighting the dependence of the error on the 10m wind speed regarding its direction.

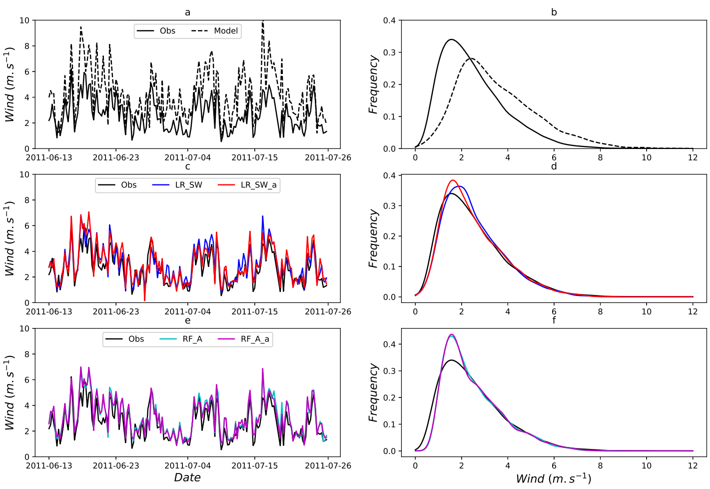

Figure 6: Timeseries (left) and PDF (right) of the observed 10m wind speed (straight black line), and 10m wind speed from ECMWF (dotted black line) (a and b), Linear models (LR_{SW} (blue) and LR_{SW}^{dir} (red)) (c and d), Random Forest models (RF_A (cyan) and RF_A^{dir} (magenta)) (e and f).

By fitting a linear model in each direction (noted with 'dir') we manually introduce a relevant information, especially for 10m wind speed (Figure 3). The model is however more exposed to under-fitting as the sample size of the training data in one direction can be low. Nevertheless, LR_{SW}^{dir} performs better than all other models. Indeed, stepwise choice is made for each direction so that the model is deeply adapted to each direction. This method results in a significant improvement of the RMSE and correlation scores. Fitting a Random Forest in each direction does not improve results, maybe because the direction is already well handled by this model by using the zonal and meridional component of the wind. So one big advantage of Random Forests over linear regression is that it does not require to explore previously deeply the data for extracting appropriate and relevant features as inputs to the model. Figure 6 shows time series of 10m observed wind speed, NWP model wind speed output over summer period of 2011 (panel a) and the probability density function corresponding to the entire period, 2011 to 2015 (panel b). Panels c and e show respectively time series of the reconstructed 10m wind speed by LR_{SW}^{dir} (red line) and LR_{SW} (blue line), and by RF_{A}^{dir} (magenta line) and RF_A (cyan line). Panels d and f show the corresponding probability density functions. All statistical models allows for a good bias correction. All models underestimate the small quantiles of the distribution and give a distribution very peaked around the mean observed wind speed. High percentiles are however well reconstructed. This is encouraging because this part of the distribution is important in terms of energy production. We can nevertheless expect an overestimation of the wind energy production with those models because of the underestimation of small percentiles.

100m wind speed reconstruction

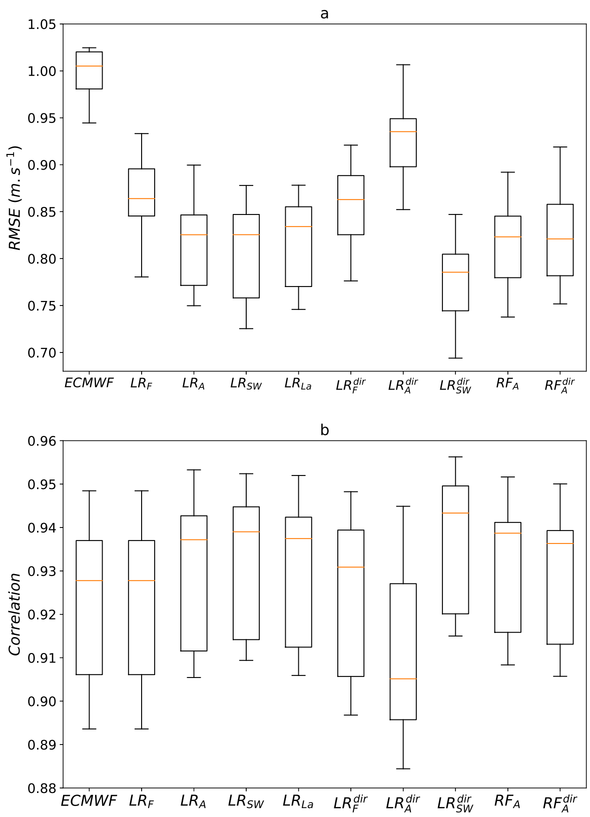

Figure 7: Same as Figure 7, for 100m wind speed.

Figure 7 show results of the reconstruction of 100m wind speed with statistical models described in Table 5. LR_F allows a reduction of the RMSE of about 15% corresponding to 0.14 $m.s^{-1}$ and the best model LR_{SW}^{dir} reduces the RMSE by 23% corresponding to 0.23 $m.s^{-1}$. The correlation is improved from 0.92 to 0.94. Adding the direction dependence to linear model with only 100m wind speed (i.e. LR_F^{dir}) does not improve results regarding LR_F. Indeed, the error on the 100m wind speed does not depend on the direction. Using all explanatory variables (i.e. LR_A^{dir}) leads to a strong overfitting. Surprisingly, the linear model using stepwise selection of explanatory variables in each direction (i.e. LR_{SW}^{dir}) recovers an important information as it performs significantly better than the other. Again, its adaptability may be the cause of its good performance. The information on the direction in Random Forests does not improve the results like for 10m wind speed reconstruction. The more important variables for Random forests and stepwise choice are mainly the 100m wind speed, but also the wind shear in the boundary layer. Lasso technique selects mainly 100m wind speed.

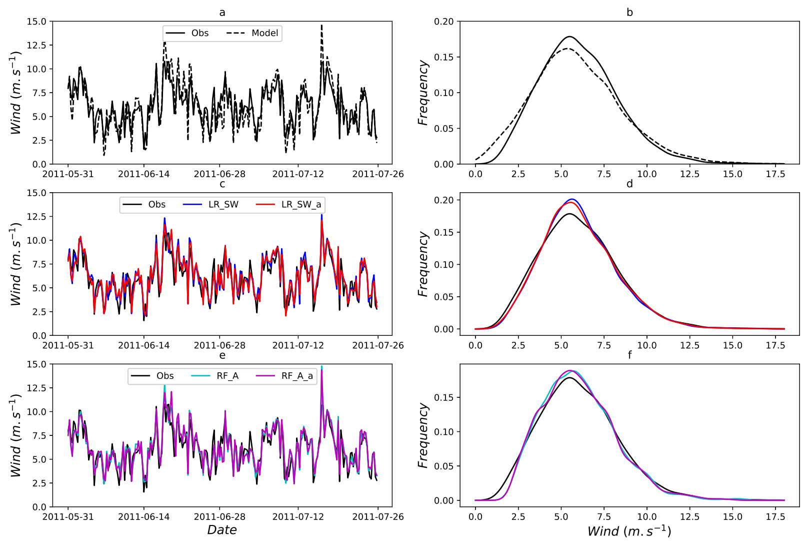

Figure 8: Same as Figure 6, for 100m wind speed.

Figure 8 shows time series of 100m observed wind speed, NWP model wind speed output over summer period of 2011 (panel a) and the probability density function corresponding to the entire period from 2011 to 2015 (panel b). panel c and e show respectively time series of the reconstructed 100m wind speed by LR_{SW}^{dir} (red line) and LR_{SW} (blue line), and by RF_A^{dir} (magenta line) and RF_A (cyan line). Panels d and f show the corresponding probability density functions. Some peaked wind speeds are less overestimated after statistical downscaling. As for the 10m wind speed, statistical models underestimate the small quantiles of the distribution and give a distribution peaked around the mean observed wind speed.

To conclude, we built different statistical models to improve the representation of the 10m and 100m wind speed of the ECMWF analyses. It has been shown that the 100m wind speed in ECMWF analyses is already well represented as it displays no systematic bias and a good correlation. Nevertheless the RMSE computed for the period 2011 to 2015 is still of 1.0 $m.s^{-1}$. Statistical models reduces the RMSE on the 10m wind speed between 40% and 65%, and between 15% and 23% for the 100m wind speed. They improve at the same time the correlation between the observed wind speed and the reconstructed one. For linear models, the variables selection is of great importance to avoid over-fitting, and an exploration step allows to improves results significantly. Random Forests give quite comparable results as the best linear models, without needing variable selection and a preliminary exploration of the data.

4 Forecasts of surface winds

In this section we use the previous statistical models based on the knowledge of the observed wind speed and the outputs of ECMWF analyses to forecast wind speed at five days horizon with a frequency of 3 hours. We have access to 1 year of forecasts in 2015. We train these statistical models on ECMWF analyses from 2011 to 2014, and apply the resulting model to the forecasts. Figures 9 and 10 show respectively the RMSE averaged over the 365 forecasts for the 10m and 100m wind speed. A strong diurnal cycle of the performances of both ECMWF forecasts and downscaled statistical predictions of the 10m wind speed is evidenced. This diurnal cycle seems to be observed also for 100m wind speed forecasts, but with a less important amplitude. As the dataset is trained on the ECMWF analyses, we can affirm that diurnal cycle is better represented in the ECMWF analyses than in ECMWF forecasts. This could be indeed explained by the data assimilation system that may help to correct errors coming from NWP model parameterizations.

An interesting result shown in Figure 9 is that performance of the LR_F statistical model which is comparable to linear model LR_{SW}, showing that the added value of other explanatory variables is valuable mainly for small lead-times in this case. Adding the dependence with the direction (i.e. LR^{dir}_{SW}) allows for a significant reduction of the RMSE. Random Forests RF_A shows the best performance. In addition to the simplicity to fit this model, its robustness seems to overcome linear regression models. For 100m wind speed forecasts (Figure 10), Linear models LR_F, LR_{SW}, and LR^{dir}_{SW} and Random Forest RF_A are comparable.

For both 10m and 100m wind speed forecasts, statistical downscaling models allow for significant improvements regarding ECMWF predicted wind speed, at any lead-time from 3 hours to 5 days. Training dataset on the analyses of ECMWF may not be optimal. Indeed, training a statistical model for each lead-time separately should deeply improve results. First, this could help to remove the displayed diurnal cycle, but may also let the increase in RMSE with the lead-time be less sharp.

5 Summary and concluding remarks

We have used statistical models to evaluate 10m and 100m wind speed at a given location from output of a NWP model. Comparison of the observed wind speed and ECMWF wind speed output at 10m and 100m within the 5 years of data show that ECMWF analyses well represent 100m wind speed.

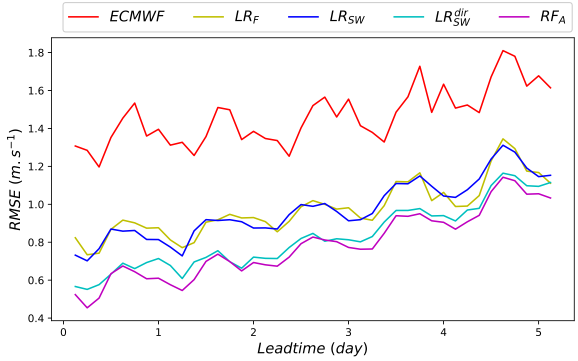

Figure 9: RMSE, computed between the 10m observed wind speed, and the 10m forecast wind speed, averaged over the entire set of forecasts.

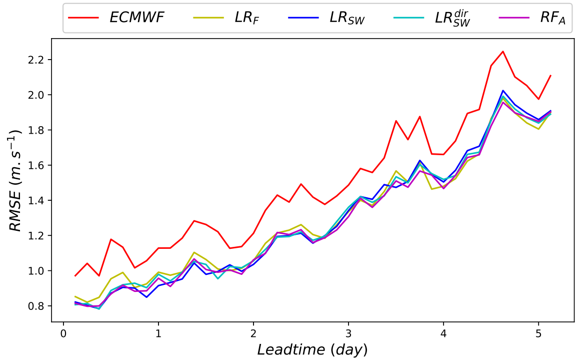

Figure 10: RMSE, computed between the 100m observed wind speed, and the 100m forecast wind speed, averaged over the entire set of forecasts.

The computed RMSE is of 1.0 $m.s^{-1}$ (the mean wind speed being of 5.8 $m.s^{-1}$) and no systematic bias is displayed. On the contrary, 10m wind speed output from ECMWF analyses displays a systematic overestimation the observed wind speed. The computed RMSE is of 1.4 $m.s^{-1}$ (the mean wind speed being of 2.4 $m.s^{-1}$).

By applying linear regression between a certain amount of selected variables and observed wind speed, we reduce the RMSE for the 10m and 100m reconstructed wind speed up to 65% and 23%, respectively. Those good results have been achieved by fitting a linear model in 8 directions and by automatic selection of valuable variables in those directions. Building such a model thus requires a special treatment and a good knowledge of the specific site so that it cannot be systematically applied to another site. Very interestingly, using Random Forests to reconstruct 10m and 100m wind speed gives comparable results as the best linear models (about 57% and 20%, respectively), while their performance is not sensitive to any preparation of the data. Computing time is a bit longer than simple linear models, but it is quite similar when a linear model is fitted in each direction.

In a second step, we applied the statistical models to forecast up to 5 days. Note that statistical models are trained on past analyses. Applying it on forecasts will work 'as well' only if the relationship between NWP outputs and observed wind speed does not change with lead-time. This is not a-priori guaranteed as the analyses incorporate information from observation via data assimilation. Results are encouraging, because the RMSE between forecast wind speed and observed wind speed is significantly reduced compared to ECMWF forecasts whatever the lead-time, and for both 10m and 100m wind speeds. Interestingly, Random Forests perform the same or better than linear models when applied to forecast 10m or 100m wind speed.

The results obtained for the forecasts are very encouraging: even though the training only involved analyses, the reduction in RMSE persisted for lead-times up to 5 days. Promisingly, there are evident changes to be tried which should lead to improvements of the performances. As a first, training statistical downscaling models directly on ECMWF forecasts makes sense as a transfer function adapted to each lead-time should take into account the reduced performance of ECMWF forecasts around 15pm and thus correct systematic errors in the representation of the diurnal cycle. Moreover, training dataset for each lead-time separately should also help to reduce the increase of RMSE with lead-time by adapting the explanatory variables to forecasts performance. For instance, for short lead-time, statistical models may find out that surface wind speed in ECMWF forecasts gather valuable information so that this information would be used. It may nevertheless not be the case at longer timescales, so that statistical models would prefer using information from large-scale circulation (e.g pressure) which is well modeled by ECMWF forecasts, even at lead-time up to 5 days. Secondly, the good performance of Random Forests together with linear regression models denotes that it is possible to reconstruct the wind speed with very different relations. Model aggregation may thus be a path to retrieve more information than when using a single statistical model.

In this study, we choose to use only local informations coming from NWP outputs. Additive valuable informations may be retrieved from larger-scale NWP outputs such as large-scale horizontal gradients of the pressure. However, the discussion on the added value of any other NWP outputs is site dependent, and is already part of research matters. For instance, it has been proved that large scale circulation patterns give valuable information at timescales up to months in some regions of France [Alonzo et al., 2017].

A wind farm is often equipped with many anemometers situated at 10m and at the hub height, so that local intime observations are easily available as well as wind power output. Forecasting wind speed using only NWP outputs is a good way to improve forecasts, but local past observations may also be used as explanatory variable. Indeed, at very short lead-time (3-hours), we can assume that the error the NWP model make at t_{0h} (corresponding to the analyses) may be correlated to the future error at time t_{3h}. We could also imagine to create a direct link between NWP outputs and wind energy production as in [Giorgi et al., 2011], using in addition the information on the close past wind energy production at the condsidered wind farm.

As a preliminary illustration of the benefit of such an approach, we train Random Forests and Linear Regression with stepwise selection of variables to forecast 10m and 100m wind speed at time t_{3h} only, and add the error

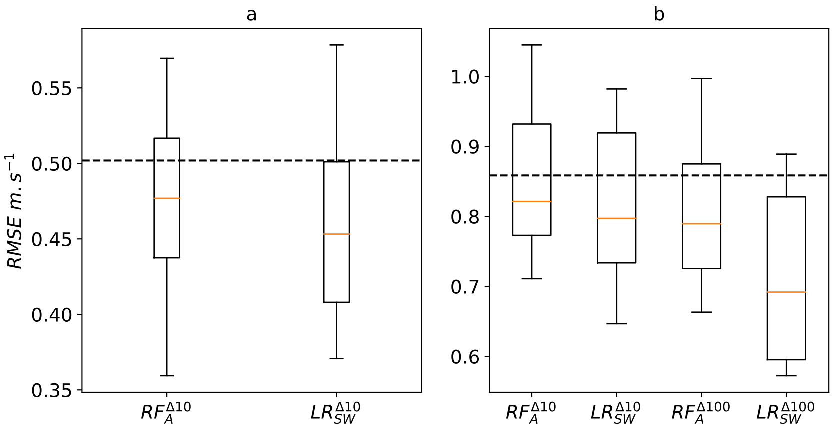

Figure 11: RMSE computed over 10 k-fold forecasts of 10m (a) and 100m (b) wind speed at 3 hour lead-time, using the error on the 10m and 100m wind speed at time t_{0h} (denoted by $\Delta 10$ and $\Delta 100$, respectively) as an explanatory variable. The dashed line represent the averaged RMSE of Random Forest without using observations at t_{0h}, and boxes represents the RMSE over 10 k-fold forecasts.

on the wind speed at time t_{0h} as an explanatory variable of the future wind speed at time t_{3h}. We perform a k-fold of 10 forecasts over the year 2015. Results are represented in Figure 11. When forecasting 10m wind speed at t_{3h}, using the error at time t_{0h} allows for a reduction of the RMSE of 5% with Random Forests and of 10% with linear model compared to Random Forest without the observation at time t_{0h}. When forecasting 100m wind speed at t_{3h}, using the information on the 10m wind speed observed at t_{0h} allows for an improvement of 2% to 6%. Adding the information on the 100m wind speed at time t_{0h} spectacularly improves results by 18% with linear regression model.

In addition of the good results obtained when reconstructing 10m and 100m wind speed, we also showed encouraging results when forecasting wind speed up to 5 days. By using very different statistical models, we highlight advantages of Random Forests over multiple linear regression, e.g simplicity and robustness. Finally, very promising perspective for improving downscaling at short-term horizon is exposed ; it involves a pseudo-assimilation of a local past observed information into the statistical downscaling model.

Acknowledgements

This research was supported by the ANR project FOREWER (ANR-14-538 CE05- 0028). This work also contributes to TREND-X program on energy transition at Ecole Polytechnique. The authors thank Côme De Lassus Saint Geniès and Medhi Kechiar who produced preliminary investigations for this study.

References

[RTE, 2016] (2016). Annual electricity report. Technical report, Reseau de Transport d'Electricite.

[GWE, 2016] (2016). Global wind statistics. Technical report, Global Wind Energy Council.

[Alonzo et al., 2017] Alonzo, B., Ringkjob, H., Jourdier, B., , Drobinski, P., , Plougonven, R., , and P.Tankov (2017). Modelling the variability of the wind energy resource on monthly and seasonal timescales. *Renewable Energy*, 113:1434–1446.

[Bauer et al., 2015] Bauer, P., Thorpe, A., and Brunet, G. (2015). "the quiet revolution of numerical weather prediction". *Nature*, 525:47–55.

[Breiman, 1996] Breiman, L. (1996). Bagging predictors. In *Machine Learning*, pages 123–140.

[Cassola and Burlando, 2012] Cassola, F. and Burlando, M. (2012). "wind speed and wind energy forecast through kalman filtering of numerical weather prediction model output". *Applied Energy*, 99:154–166.

[Davy et al., 2010] Davy, R. J., Woods, M. J., Russell, C. J., and Coppin, P. A. (2010). Statistical downscaling of wind variability from meteorological fields. *Boundary Layer Meteorol*, 175:161–175.

[Devis et al., 2013] Devis, A., van Lipzig, N., and Demuzere, M. (2013). A new statistical approach to downscale wind speed distribution at a site in northern europe. *Journal of geophysical research : Atmospheres*, 118:2272–2283.

[Folland et al., 2012] Folland, C. K., Scaife, A. A., Lindesay, J., and Stephenson, D. B. (2012). "how potentially predictable is northern european winter climate a season ahead?". *Int. J. Climatol.*, 32:801–818.

[Friedman et al., 2001] Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

[Giorgi et al., 2011] Giorgi, M. G. D., Ficarella, A., and Tarantino, M. (2011). "assessment of the benets of numerical weather predictions in wind power forecasting based on statistical methods". *Energy*, 36:3968–3978.

[Glahn and Lowry, 1972] Glahn, H. R. and Lowry, D. A. (1972). "the use of model output statistics (mos) in objective weather forecasting". *Journal of Applied Meteorology*, 11:1203–1211.

[Haeffelin et al., 2005] Haeffelin, M., Barthes, L., Bock, O., Boitel, C., Bony, S., Bouniol, D., Chepfer, H., Chiriaco, M., Cuesta, J., Delanoe, J., Drobinski, P., Dufresne, J.-L., Flamant, C., , Grall, M., Hodzic, A., Hourdin, F., Lapouge, F., Lemaitre, Y., Mathieu, A., Morille, Y., Naud, C., Noel, V., OHirok, W., Pelon, J., Pietras, C., Protat, A., Romand, B., Scialom, G., and Vautard, R. (2005). "Sirta, a ground-based atmospheric observatory for cloud and aerosol research". *Annales Geophysicae*, 23:1–23.

[Howard and Clark, 2007] Howard, T. and Clark, P. (2007). "correction and downscaling of nwp wind speed forecasts". *Meteorol. Appl.*, 14:105–116.

[Jung and Broadwater, 2014] Jung, J. and Broadwater, R. P. (2014). "current status and future advances for wind speed and powerforecasting". *Renewable and Sustainable Energy Reviews*, 31:762–777.

[Kalnay, 2003] Kalnay, E. (2003). *Atmospheric Modeling, Data Assimilation and Predictability.* Cambridge University Press.

[Kubik et al., 2011] Kubik, M. L., Coker, P. J., and Hunt, C. (2011). Using meteorological wind data to estimate turbine generation output: a sensitivity analysis. In *Proceedings of Renewable energy congress*, pages 4074–4081.

[Maraun et al., 2010] Maraun, D., Wetterhall, F., Ireson, M., Chandler, E., Kendon, J., Widmann, M., Brienen, S., Rust, H. W., Sauter, T., Themel, M., Venema, V. K. C., Chun, K. P., Goodess, C. M., Jones, R. G., Onof, C., Vrac, M., and ThieleEich, I. (2010). "precipitation downscaling under climate change : Recent developments to bridge the gap between dynamical models and the end user.". *Review of Geophysics*, 48:RG3003.

[Mohandes et al., 2011] Mohandes, M., Rehmanb, S., and Rahmand, S. (2011). "estimation of wind speed prole using adaptive neuro-fuzzy inference system (anfis)". *Applied Energy*, 88:4024–4032.

[Salameh et al., 2009] Salameh, T., Drobinski, P., Vrac, M., and Naveau, P. (2009). "statistical downscaling of near-surface wind over complex terrain in southern france". *Meteorol Atmos Phys*, 103:253–265.

[Soman et al., 2010] Soman, S. S., Zareipour, H., and O. Malik, P. M. (2010). "a review of wind power and wind speed forecasting methods with different time horizons". *North American Power Symposium (NAPS)*, pages 1–8.

[Tibshirani, 1994] Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

[Vallis, 2006] Vallis, G. K. (2006). *Atmospheric and Oceanic Fluid Dynamics.* Cambridge University Press, Cambridge, U.K.

[Wagenbrenner et al., 2016] Wagenbrenner, N. S., Forthofer, J. M., Lamb, B. K., Shannon, K. S., and Butler, B. W. (2016). "downscaling surface wind predictions from numerical weather prediction models in complex terrain with windninja". *Atmos. Chem. Phys.*, 16:5229–5241.

[Wilby and Dawson, 2013] Wilby, R. L. and Dawson, C. W. (2013). "the statistical downscaling model: insights from one decade of application". *Int. J. Climatol.*, 33:1707–1719.

[Wilby et al., 1998] Wilby, R. L., Wigley, T. M. L., Conway, D., Jones, P. D., Hewitson, B., Main, J., and Wilks, D. S. (1998). "statistical downscaling of general circulation model output: A comparison of methods". *Water Resources Research*, 34:2995–3008.

[Zamo et al., 2016] Zamo, M., Bel, L., Mestre, O., and Stein, J. (2016). "improved gridded wind speed forecasts by statistical postprocessing of numerical models with block regression.". *Weather and Forecasting*, 31:1929–1945.