Wind Power forecasting for the day-ahead energy market

Provider : CNR (Compagnie Nationale du Rhône)

Context

CNR is the French leading producer of exclusively renewable energy (water, wind, sun) and the concessionary of the Rhone river for hydroelectricity production, river navigation and irrigation for agricultural use.

This challenge focuses on wind energy production forecast. CNR currently owns around 50 Wind Farms (WF) for a total installed capacity of more than 600 MW. Every day, CNR sells on the energy market its wind energy production for the day ahead. In order to sell the right amount of energy, as well as for legal requirements towards the French Transmission System Operator (TSO) in charge of the electric network stability, CNR needs to know beforehand how much energy the wind farms will produce the day ahead.

Goal of the challenge

The goal of this challenge is to predict the energy production of six WF owned by CNR. Each WF production will be individually predicted, as well as the energy production of the whole portfolio represented by the sum of the six WF individual energy production. Predictions will focus on the day-ahead energy production (hourly production forecasts from day D+1 00h to day D+2 00h).

Problem description

The competitors will have access to the six WF hourly production data from May the 1st of 2018 to January the 15th of 2019 (8 months and 15 days). This defines the training dataset, since day-ahead WF hourly production is the prediction target (predictand). Provided hourly WF power production consists in the raw recordings of the DSO (Distribution System Operator), and should therefore be considered as a reference, even if it could contain erroneous or suspect data. The choice is left to competitors to criticize or not this data, using complementary data provided apart.

The test period will run from January the 16th of 2019 to September the 30rd of 2019 (8 months and 15 days). Evaluation will be performed on raw observed hourly WF power production.

For confidentiality purposes, the names of the six considered WF will not appear in the challenge, and we will use generic names instead: WF1, … , WF6. For similar reasons, locations of the WF will not be provided.

For both training and test periods, predicting variables (predictors) will be given. It consists in hourly forecasted meteorological variables, provided by various Numerical Weather Prediction (NWP) models. NWP models are meteorological models run by several national weather prediction services.

For confidentiality reasons, the name of the NWP models will not appear. They will be named with generic names NWP1, … , NWPn.

We propose to provide also complementary observed data, which could not be used as predicting variables (because these data are not subject to forecasts), but which may help competitors to prepare or criticize WF hourly production data. This consists in hourly wind speeds and wind directions observed at the height of each wind turbine, on each WF. Obviously, complementary data will not be provided for the test period.

Important remark on the evaluation: in order to simulate real operational forecasting conditions, we impose that day-ahead predictions must be computed with data that are available in real operational conditions on day D at 09h00 UTC. Consequently, only these data will be provided in the test set, whereas a much larger set of data will be provided in the training set. On this topic, we strongly recommend the competitors to refer to the document named "formatting_Xtrain_Xtest_files.xlsx" in the complementary data zip archive.

Data description

The format of the train and test .csv file is the following:

ID	WF	Time (UTC)	Variables from NWP1			…	Variables from NWP N		
			Run 1 - Variable 1	…	Run M - Variable L	…	Run 1 - Variable 1	…	Run M - Variable L
1	WF1	DD/MM/YYYY HH:MM				….			
2	WF1								
…	…								
X	WF6								

ID	Observed Power Production
1	
2	
…	
X	

Here is a description of all data provided in the input csv files:

ID: This is the unique ID of each row in the csv files. One ID correspond to a couple Time / WF. The ID of the test set are consecutive to the ID of the training set.

WF: The considered Wind Farm. WF ranges from WF1 to WF6. It is crucial for the competitors to be aware that this prediction problem is totally dependent to the WF considered. In other words, the statistical link between input variables and wind power production is completely different from one WF to another. Consequently, it could be judicious to train specific prediction algorithms for each WF, instead of training a unique algorithm which could be unable to model the behavior of each WF.

Time (UTC): date and hour of the target timestep, i.e. corresponding to the observed Power production. Time zone is Coordinated Universal Time (UTC).

Meteorological variables: Numerical Weather Predictions are provided by meteorological centers several times a day (updates), typically at 00h UTC, 06h UTC, 12h UTC and 18h UTC. We call these sets of forecasts "Runs". Consequently, if the input file contains forecasts arising from several runs, this implies that a single NWP is associated with several forecasts for the same forecasting time. There|fore, the information on the hour of run is provided.

The format of the header of the csv files for the meteorological variables is the following:

NWPi_HourOfTheRun_DayOfTheRun_Variable

With *NWPi* the considered Numerical Weather Prediction model (meteorological model);

HourOfTheRun the hour (UTC) of the considered run. According to the NWP, it could be 00h, 06h, 12h and 18h (case of NWP with 4 runs per day) or only 00h and 12h (case of NWP with 2 runs per day);

DayOfTheRun the day of the considered run. We provide in the csv files predictions from the D-2 day runs (the day before yesterday), D-1 day runs (yesterday) and D day runs;

Variables the different meteorological variables forecasted by the NWP. These are essentially U, V and T:

- *U and V components of the wind at 100m (or10m) height (m/s)*: these are the zonal and meridional velocities of the wind, respectively. Both are given at a height of 100m above ground for NWP1, NWP2 and NWP3. U and V are given at a height of 10m for NWP4. Even if these variables are given at hourly timestep, we draw competitors attention on the fact that the temporal representativity of the given values is for a 10-minutes window ranging from H-10 min to H.

  Additional remark: since wind power production is principally driven by the wind speed impacting turbines, it could be useful for the competitors to derive wind speed (and wind direction) from U and V. This can be done using a simple trigonometric calculation of the magnitude and direction of a vector with U and V components. The choice is let to the competitors.

- *Temperature of air (°C), abbreviated T:* this is the averaged temperature over the entire hour (from H-1 to H). Wind power production is sensitive to air temperature since it affects the air density. This variable is provided only for NWP1 and NWP3.

- *Total cloud cover (%), abbreviated CLCT:* this is the total cloud cover of the sky, ranging from 0% (clear sky, no cloud) to 100% (fully clouded sky). The value is an instant value at hour H. This variable is provided only for NWP4.

- *Other variables:* We may provide other meteorological variables in the future, before the start of the challenge. If so, we will provide an update of this technical description, including the definition of these new variables.

Important remark 1: Some of the NWP provide hourly forecasts, which means that predicted variables are available on every timestep in the csv files, but others provide only three hourly forecasts. For the corresponding columns, blanks are let for timesteps that are not subject to forecasts. The choice is let to competitors to reconstruct or not the missing values to get hourly time series for all the NWP.

Important remark 2: As stated before, day-ahead predictions must be computed with data that are available on day D at 09h00 UTC, and data provided in test sets will be restricted in that way. Practically, this means that at 09hUTC competitors will have access to 00h UTC runs only (and to runs from the D-1 day), because 06hUTC runs are not available at this time. This is an important remark to account for in the construction of the methodology, even if this does not mean that the inputs unavailable in the test set are not usefull for the training.

Observed Power Production (MW or MW.h): this is the observed total amount of energy injected by the WF to the electric network over the entire hour H-1 to H (MW.h). Equivalently, we can consider that this is the mean power output of the WF observed between H-1 and H (MW).

WF complementary data

We provide complementary data in the .zip supplementary files. These data may be used by the competitors to prepare or criticize WF hourly production data, but they are not predictors. The file WindFarms_complementary_data.csv contains the following hourly variables:

- *Average power output for each wind turbine of the WF (MW)*
- *Cumulated energy produced by each wind turbine (MWh). This value could differ from the hourly average power output when the considered turbine has not been operational during the entire hour.*
- *Observed average wind direction at hub (nacelle) height for each wind turbine (°, from 0 to 359)*
- *Observed average wind speed at hub (nacelle) height for each wind turbine (m/s)*
- *Observed average nacelle direction for each wind turbine (°, from 0 to 359)*
- *Observed average rotational speed of each wind turbine (s^{-1})*

Metric

The metric used to rank the predicting performance is a relative form of the absolute error. We call it the CAPE (Cumulated Absolute Percentage Error). The formulation of CAPE for one WF would be the following:

$$CAPE_k\left(\hat{Y}_k, Y_k\right) = 100 \times \frac{\sum_{i=1}^{N_k}\left|Y_{i,k} - \hat{Y}_{i,k}\right|}{\sum_{i=1}^{N_k} Y_{i,k}}$$

With $CAPE_k$ the metric for the WF k (%)

N_k the length of the test sample for WF k only

$Y_{i,k}$ the observed production for WF k and hour i (MW or MW.h)

$\hat{Y}_{i,k}$ the predicted production for WF k and hour i (MW or MW.h)

For convenience reasons, data relative to the 6 WF have been regrouped in the same train and test input files. Therefore, the metric used in the challenge is the overall average CAPE for the 6 WF, calculated as:

$$CAPE\left(\hat{Y}, Y\right) = 100 \times \frac{\sum_{i=1}^{M}\left|Y_i - \hat{Y}_i\right|}{\sum_{i=1}^{M} Y_i}$$

With M the length of the test sample for all the 6 WF (M is the sum of N_k for all k)

This formulation results in a non-homogeneous contribution of all the WF to the final value of CAPE: CAPE will be more sensitive to WF with the highest energy production values.

Remark relative to the separation of data into training / public test / private test sets:

Due to the use of a chronological sampling to separate training / public / private test sets, seasonal effects probably play an important role on the results (metric) obtained for each set of data, because of the non-homogeneous distribution of wind energy production values for these three sets. This means that competitors top-ranked on the public test set will not necessarily be ranked among the bests on the private test set. We are aware of this weakness in our evaluation protocol, but we could not do differently without disturbing the temporal continuity of data in a harmful way. In the competitors' point of view, this is an important remark to account for.