Gradient boosting-based approach for short- and medium-term wind turbine output power prediction

Robert Adam Sobolewski
Bialystok University of Technology, Faculty of Electrical Engineering, 15-351 Bialystok, Wiejska 45D street, Poland, e-mail: r.sobolewski@pb.edu.pl

Médane Tchakorom
Univ. Bourgogne Franche-Comte (UBFC), FEMTO-ST Institute, CNRS, Belfort, France, e-mail: medane.tchakorom@univ-fcomte.fr

Raphaël Couturier
Univ. Bourgogne Franche-Comte (UBFC), FEMTO-ST Institute, CNRS, Belfort, France, e-mail: raphael.couturier@univ-fcomte.fr

Abstract: Being able to predict the output power of wind turbines and wind farms is crucial in the process of integrating such stochastic energy sources with power systems. To support stakeholders in short- and mid-term wind power prediction, a novel data-driven Machine Learning based approach is proposed. This approach relies on three Gradient Boosting (GB) regressor implementations. The novelty of our approach is also manifested in the fact, that it is respectively based on the use of MERRA-2 reanalysis data and GEOS FP meteorological forecasts in models training and wind power prediction. It makes the short- and mid-term prediction unique in enriching the results even for time horizon of 240 h with resolution of 1 hour. The data preprocessing and cleaning, feature engineering, and training, testing and validation of the models are presented in details. The performances of the models and prediction accuracy are evaluated relying on a few absolute and relative error measures. The proposed methodology is implemented in the output power prediction of a wind turbine located in Poland. The results of predictions are compared with other Machine Learning algorithms. The results show that proposed GB implementations can capture accepted accuracy of prediction and outperform other investigated algorithms.

Keywords: wind power, short-and medium-term prediction, gradient boosting, MERRA-2, GEOS FP

1. Introduction

Although the success of wind energy is increasing significantly year by year, it still presents essential operational and planning challenges to both wind farm operators and power system operators. The challenges are caused by the stochastic nature of wind resources and other meteorological features, such as wind dynamics, wind turbines operational characteristics and many more features. An essential part of the effective integration of wind energy generation units lies in the accurate prediction of their output power. This is crucial to power system operators – for optimal load flow analysis, voltage levels control and unit commitment studies, and wind farm operators – for biding at day ahead energy market and maintenance planning. One of some reasonably options of wind turbine (or farm) output power prediction (in short – wind power prediction) is to use available historical weather data and historical wind turbine (farm) data as well as meteorological forecast.

Wind power prediction studies, are broadly classified into direct and indirect approaches [1]. Concerning direct approaches, different methods enable one to directly predict wind power. The main advantage of this approach is that there is no need to investigate the relationship between output power and meteorological features, e.g. wind speed and direction. In latter approaches a wind speed is firstly forecasted and then the resulted data is converted into output power relying on different techniques (e.g. deterministic power curve, statistical models). In practice, while transforming wind speed into output power, further errors are made in prediction accuracy. Statistical models seem to be better than power curves to incorporate the nonlinear and uncertain relationships between output power and wind speed. There are few papers in related literature that report the comparison of the performance of both direct and indirect approaches [1]. Most of them show that direct approaches offer the best predicting accuracy.

Wind power prediction models are usually categorized as physical and statistical models [2]. Both approaches are able to predict the output power effectively, but they are profoundly different in terms of the ideas that are applied. Physical models use mathematical expressions to model highly complex and nonlinear dynamics of the atmospheric flow to produce numerical weather predictions (NWP). Then NWP can be adopted to local flow conditions, focusing mainly on wind speed, which can be finally used as an input in the wind power prediction. Statistical models rely on relevant chronological data either wind power data (direct approach) or wind speed data (indirect approach). They use methods such as the Autoregressive Moving Average (ARMA) or the Autoregressive Integrated Moving Average (ARIMA). They are able to forecast the variables in question with one or a few time steps ahead. Another classification of the approaches of wind power prediction is their division into: (i) model-driven, (ii) data-driven, and (iii) hybrid intelligent approaches [3]. In the model-driven approaches, abundant meteorological information of distinct physical factors affecting wind power is required. NWP model is an example of this approach. In data-driven approach, statistical modelling based on historical data are used in the predictions. The ARMA and ARIMA models are examples of data-driven models. Another group consists of Machine Learning (ML) algorithms, e.g. Gradient Boosting (GB), Artificial Neural Networks (ANN), Decision Trees (DT), Random Forests (RF), and many more. Since wind power (and wind speed) time-series is highly stochastic in nature, the prediction accuracy relying on data-driven approaches may be unsatisfactory. Such problems can be overcome with the hybridization of two or even more either data-driven [1, 3, 4, 5] or model-driven and data-driven methods [6, 7]. Paper [7] presents a summary of the recently proposed fifteen wind power prediction models using both statistical data-driven and hybrid methods. The models that require either one input data (wind power or wind speed) or several input data (combinations of either wind power and meteorological features or meteorological features only). The time horizon of prediction ranges from 30 min up to 250 h (usually a few hours only, sometimes dozens of hours). A comprehensive review of the various deep learning algorithms being used in wind power (and wind speed) forecasting including the stages of data processing, future extraction and learning issues is presented in [8].

Both data-driven approaches and hybrid intelligent models can be trained relying on historical data (one or more input features) to predict target feature (wind power). If historical data concerns one input feature it can be either output power or wind speed, depending on the modelling purpose – direct or indirect prediction. In the case of more than one input feature as a historical data, they can be both output power and meteorological features that most affect the output power, and the target feature. In the latter case of direct power prediction its accuracy can improve significantly when additional meteorological forecasts are used. The only condition is full compatibility between the historical meteorological data and weather forecasts in terms of: type of features, geographical coordinates, and resolution.

Unfortunately, meeting the condition of access to historical meteorological data and relevant weather forecasts is difficult or impossible for most wind turbines, farm operators, or grid operators. This problem is more evident while predicting output power of offshore wind farms. From this point of view, it is useful to use climate reanalysis data and weather forecasts that are compatible with them, for short- and mid-term output power predictions (time horizon from one up to tens or several hundred hours). Many reanalysis products are available [9, 10]. Wind speeds and direction are most commonly available at a fixed height of 10 m above ground (common met mast altitude). A key benefit of reanalysis is that it can infer meteorological features for which there are no observations, i.e. in locations that are either remote or out at sea (where met masts are not present). Data from climate reanalyses are useful data in wind power analysis, with large spatial and temporal coverage.

The use of second Modern-Era Retrospective Analysis for Research and Applications (MERRA-2) reanalysis and of the Goddard Earth Observing System – Forward Processing (GEOS FP) which is compatible with it can be very beneficial to wind power prediction. MERRA-2 reanalysis [11] is one of the major state-of-the-art climate reanalysis freely available on a global scale. It covers many meteorological features that significantly affect wind power investigations. The advantage of this reanalysis is other model heights, usually based on fixed pressure isothermal level. For example wind speed and direction at 50 m (met masts are usually only 10 m tall). It provides multiple meteorological features at the hourly time step, with a spatial resolution of ~50 km (corresponds to the resolution of 0.625^{o} longitude by 0.5^{o} latitude) and temporal coverage from 1980 onwards. In addition, MERRA-2 reanalysis is published with a short latency of two or three weeks of real time, which makes it suitable for updating the predicting model. GEOS FP weather forecasts [12] with high spatial resolution 0.3125^{o} longitude by 0.25^{o} latitude were adopted from GEOS-5 of the Data Assimilation System (DAS). This is a resource widely used and validated in studies of atmospheric chemistry that combines a suite of observations, including data from satellites, radiosondes, aircrafts, dropsondes, surface ships, and buoys. As regards historical wind turbine or wind farm data, the data can be retrieved from built-in Supervisory Control and Data Acquisition (SCADA) systems. These data both refer to the operational data (measurements of electrical, mechanical and other quantities) and to the operational statuses of the turbines.

The main objective of the study is to explore the use of state-of-the-art ML techniques to formulate new GB models to predict wind turbine output power relying on both historical turbine data collected by SCADA system and on meteorological data gathered from MERRA-2 reanalysis, and relying also on weather forecasts being retrieved from GEOS FP. Three implementations of GB are investigated, i.e. CatBoost Regressor, LightBoost Regressor and XGBoost Regressor. The study applies advanced data preprocessing, feature engineering, GB models formulation, their parameters tuning to deliver improvements in terms of prediction accuracy, generalization ability as well as computational performance for wind turbine output power prediction. Their validity was comprehensively assessed by comparing the LSTM model, the Decision Tree model and the Random Forest model. In [13], the authors compared the GB and LSTM, for times series predictions. There are not many examples in the literature yet of using GB for wind power prediction. In [14] a hybrid model of wind power prediction based on XGBoost and NWPs was presented. The same GB implementation together with weather similarity analysis and feature engineering was applied in [15] for short-term wind power forecasting.

GB were proven to be superior to traditional ANNs which implies a great potential for wind power prediction. The proposed GB models are suitable to predict the output power of wind turbines installed both onshore and offshore, in any location (country, continent). In the case of offshore turbine locations, the presented approach can be exceptionally competitive

because it is not possible to obtain actual measurements of meteorological features, as such measurements are not carried out. Other advantages of the GB models over the methodologies presented in related literature are as follows: (i) there is no need to convert meteorological values (especially wind speed and direction) to the height corresponding to the height of the turbine nacelle above the ground surface, and (ii) there is no need to convert wind speed into turbine power according to the power curve provided by the manufacturer or any statistical models. Due to the use of both meteorological databases (reanalysis and weather forecasts), there is no need to use historical data covering the time period up to *t* in order to make forecasts starting from time *t*. Thus, the last historical data may come from days or months ago, and the prediction can be made from the current time.

The novelty and scientific originality of this paper is that it is the first reported study which implements GB algorithm for wind power prediction where key input parameters are gathered from SCADA system of wind turbine, MERRA-2 reanalysis and GEOS FP meteorological forecasts. The use of MERRA-2 and GEOS FP allows to wind power prediction in a time horizon of up to 240 h with resolution of 1 h. Moreover, there is need to convert any meteorological features (mainly wind speed) into wind turbine output power and to use historical data covering the time period up to *t* in order to make prediction starting from time *t*. The novelty and main contributions of the work can be summarized as follows:

a) data cleaning method was introduced to effectively detect outliers in original wind turbine output power data to improve the quality of input data; it relies on expert knowledge and operational statuses recorded by the SCADA system of the turbine,
b) the methodology uses many meteorological features provided by MERRA-2 reanalysis and GEOS FP forecasts, i.e. surface pressure, 2- and 10-meter m specific humidity, 2- and 10-meter air temperature, and 2-, 10- and 50-meter wind speed and direction; a new idea to use these features from four nearest grid points around the wind turbine site coordinates is proposed; this allows for wind power predicting regardless of the turbine site (onshore and offshore) and availability of local meteorological data measured thanks to meteorological masts,
c) advanced feature engineering was introduced to improve the accuracy of prediction; it refers to categorical and cyclic features, and data normalization; original wind direction is additionally transformed into categorical feature, i.e. 16 cardinal (and intercardinal) wind directions; cyclic features (i.e. hours of day, days of week, months of year, and wind direction) are transformed based on sin and cos trigonometric functions; two calculated features substitute each original cyclic feature,
d) the GB algorithm and its three implementations CatBoost, LightBoost and XGBoost are proposed for effective and accurate solving the regression problem among input features and target feature (output power) using the ‘boosting trees’ approaches,
e) the GB based methodology has been validated using historical data from E-53 wind turbine by Enercon, operating in north-eastern Poland; the accuracy of wind power prediction relying on three GB implementations is better as compared to the accuracy of prediction based on other ML algorithms (e.g. LSTM neural networks, Decision tree and Random Forest); moreover, the approach generalizes and significantly improves the accuracy of wind power prediction as compared to other methodologies that rely on time series prediction; with the extension of the predicting time horizon, this advantage over other methods becomes more and more evident.

The remainder of the paper is organized as follows. Section 2 provides some details on four machine learning algorithms, i.e. GB, LSTM, Decision Tree and Random Forest. Concerning GB, three implementations are presented – CatBoost, XGBoost and LightBoost. Section 3 introduces the idea and detailed description of the wind turbine output power

prediction process which covers the following steps: obtaining the required data, data pre-processing and features engineering, training the prediction models, their hyper-parameters tuning, and finally – obtaining some power predictions. Section 4 presents case study of real wind turbine output power prediction and discussion of the results. Section 5 concludes the study by summarizing the key findings and the contributions of this work.

2. Machine learning algorithms

2.1. GB algorithms

Boosting algorithms refer to a class of learning algorithms that fit models by combining many sampler models [16]. These sampler models are typically base models and are trained using base learner or weak learner. The models tend to perform slightly better than a random guess, but when selected carefully and aggregated using a boosting algorithm, they form a relatively more accurate model. In our work, three boosting implementations were used, i.e. CatBoost, LightBoost and XGBoost. To be more precise, gradient boosting regressors are used to solve the regression problem using the 'boosting trees' approaches.

A benefit of using gradient boosting is that once the boosted trees are computed, it is quite straightforward to retrieve importance scores for each feature (attribute). Generally, importance provides a score that indicates how useful or valuable each feature was in formalizing the boosted decision trees within the model. The more a feature is used to make key decisions with decision trees, the higher its relative importance. This importance is calculated explicitly for each feature in the dataset, allowing attributes to be ranked and compared to each other. Importance is calculated for a single decision tree by the amount that each feature split point improves the performance measure, weighted by the number of observations the node is responsible for. The performance measure may be the purity score used to select the split points or another more specific error function. The feature importance are then averaged across all of the decision trees within the model.

Catboost regressor implementation

Catboost is a very efficient GPU and CPU implementation of GB. Despite being launched later than other well-known boosting algorithms, Catboost has developed a high reputation for its speed of execution and the accuracy of its predictions. It was open sourced by yandex, one of Russia's largest internet businesses, in April 2017. It distinguishes itself from its competitors by a number of qualities, including:

- ordered Boosting to overcome over fitting,
- native handling for categorical features,
- using Oblivious Trees or Symmetric Trees for faster execution.

Lightboost regressor implementation

By using a leaf-wise tree growth method and introducing unique techniques Light Gradient Boosting Machine (LightGBM), a recent upgrade of the gradient boosting algorithm, was able to maintain its excellent predictivity while also resolving its scalability and long processing time. Microsoft announced the first stable version of LightGBM in January 2017. This implementation of the GB algorithm is known for its speed. Compared to its competitors, LightGBM has the following features:

- leaf wise tree growth compared to other boosting algorithms that use level-wise tree growth,

- faster training speed,
- lower memory usage.

XGboost regressor implementation

One of the most well-known implementations of Boosting algorithms is eXtreme Gradient Boosting (XGBoost) [17]. It began as a Tianqi Chen research project and has since grown in popularity in the machine learning competition circles. XGBoost is an open-source toolkit that is part of the Distributed Machine Learning Community (DMLC). As such, it benefits from a strong and active community and is more transparent than its competitors, allowing for easy plotting of trees, for example. Other characteristics that distinguish XGBoost from other boosting solutions include:

- a proportional shrinking of leaf nodes,
- clever penalization of trees,
- Newton Boosting,
- extra randomization parameter.

2.2. ANN algorithms

ANNs emerged as one of the most commonly used ML algorithms in the field of wind power prediction [2]. ANNs are complex structures that attempt to mimic the structure of the human brain based on a set of replicated processing units called neurons. Neurons are interlinked and pass information via weighted connections adjusted during the training process. Developments in initialization algorithms and neuron activation functions enhanced the capabilities of ANN and made it possible to solve complex non-linear problems by training models consisting of a large number of hidden layers, that is often referred to as 'deep learning'. Recurrent neural network (RNN) is a class of an ANNs, in which the connection between its neurons form a loop, allowing information to persist. This means it is capable of handling nonlinear dependencies between past time series values and the estimate of values to be predicted via the inherent dynamic memory created by recurrent connections in the hidden layers. Despite its superiority over conventional ANNs [18], RNN suffers from a phenomenon referred to as vanishing or exploding gradients caused by error signals flowing backwards, which leads to oscillating weights or loss of long-term dependencies due to the rapid decay (vanishing) or increase (exploding) in the norm of gradient during training. Among the numerous methods proposed to address vanishing and exploding gradients, the introduction of gating mechanisms to control the flow of information between layers has shown promising results and practical applications. Some noteworthy examples of RNN architectures adopting this principle are Gated Recurrent Unit (GRU) and Long-Short Term Memory (LSTM) [19]. Existing studies on wind power prediction relying on ANN have mainly been based on LSTM models. The data transfer process is LSTM is similar to that of standard recurrent neural networks. However, the information propagation operation is different. As the information passes through, the operation decides what information to process further and what information to discard. The main operation consists of cells and gates. The state of the cell functions as a pathway for the transfer of information. You can think of cells as a memory. For more details about LSTM, interested readers are invited to read [19].

2.3. Decision Tree

Roots, branches, and leaves make up a typical tree. Decision Trees follow the same structure. There are root nodes, branches, and leaf nodes in it. Every internal node is used to test an attribute, the result of the test is on the branch, and the class label is on the leaf node as a result [20, 21]. A root node is the topmost node in a tree and serves as the parent of all nodes. A decision tree is a tree in which each node (attribute) represents a characteristic, each

link (branch) represents a decision (rule), and each leaf represents an outcome (categorical or continuous value) [21]. Because decision trees are designed to replicate human reasoning, grabbing facts and making good interpretations is extremely easy. The goal is to build a similar tree for all of the data and process a single result at each leaf.

2.4. Random Forest

Random Forest is an algorithm based on the assembly of decision trees. It consists of a set of independent decision trees. Each tree has a fragmented view of the problem due to a double random draw:

- a random draw with replacement of the training set; this process is called tree bagging,
- a random draw on the variables; this process is called feature sampling.

In the end, all these independent decision trees are assembled. The prediction made by the Random Forest for unknown data is then the average (or the vote, in the case of a classification problem) of all the trees.

A Random Forest algorithm can perform both regression and classification tasks. It produces efficient predictions that can be understood easily and it provides a higher level of accuracy in predicting outcomes over the decision tree algorithm.

3. MERRA-2 reanalysis and GEOS FP weather forecasts

MERRA-2 is a NASA global atmospheric reanalysis [11]. It uses Goddard Earth Observing System Model, Version 5.12.4 (GEOS-5) data assimilation system. This system combines historic weather observations with the atmospheric circulation model to infer the state of the global weather conditions. The model is set to replicate historic observations from satellites, ground observatories, ships, aircrafts, etc. producing a hindcast as opposed to forecast. Data collections from MERRA-2 are provided on the horizontal grid that has 576 points in the longitudinal direction and 361 points in the latitudinal direction, corresponding to the resolution of 0.625^{o} longitude $\times$ 0.5^{o} latitude (cubed-sphere grid with an approximate resolution of 50 km $\times$ 50 km). The collection ‘tavg1_2d_slv_Nx (M2T1NXSLV)’ [22] is one of the most interesting ones for wind power prediction among all MERRA-2 collections provided. Temporal coverage of the data collection is from 1980 onwards.

GEOS FP is a global atmospheric data assimilation system designed for analyses and forecasts produced in real time, using the most recent validated GEOS assimilation system [12]. The GEOS FP global horizontal grid consists of 1152 points in the longitudinal direction and 721 points in the latitudinal direction, corresponding to the resolution of 0.3125^{o} longitude $\times$ 0.25^{o} latitude (cubed-sphere grid with an approximate resolution of 25 km $\times$ 25 km). The collection ‘tavg1_2d_slv_Nx’ is useful in most wind power short-term prediction (max 240 hours ahead) investigations, among all GEOS FP collections provided. The dataset is uploaded every day and temporal coverage of the data is from 00:30 of the first day (the day before withdrawal of data) till 23:30 of the tenth day.

Both MERRA-2 and GEOS FP ‘tavg1_2d_slv_Nx’ collections are single-level diagnostics, of 1-hourly frequency (data averaged over 1 hour) with spatial grid 2D. One MERRA-2 file covers the data records for 24 h from 00:30 UTC till 23:30 UTC, whereas one GEOS FP file consists of one data row for 1 h (time-stamped with the central time of the interval, i.e. 00:30 UTC, 01:30 UTC, 02:30 UTC, etc.). Since the horizontal native grid origin represents a grid point located at 180^{o}W, 90^{o}S in both MERRA-2 and GEOS FP files, the points that correspond to the resolution of 0.625^{o} $\times$ 0.5^{o} are the same in both files. The list of meteorological variables useful in the process of wind power prediction provided in both

MERRA-2 and GEOS FP ‘tavg1_2d_slv_Nx’ collections is the following: DISPH – zero plane displacement height [m], PS – surface pressure [Pa], QV10M – 10-meter specific humidity [kg/kg], QV2M – 2-meter specific humidity [kg/kg], SLP – sea level pressure [Pa], T10M – 10-meter air temperature [K], T2M – 2-meter air temperature [K], U10M – 10-meter eastward wind [m/s], U2M – 2-meter eastward wind [m/s], U50M – eastward wind at 50 meters [m/s], V10M – 10-meter northward wind [m/s], V2M – 2-meter northward wind [m/s], and V50M – northward wind at 50 meters [m/s]. Additional details about variables listed in the files specification [11, 12] can be found in [23].

The provider of GEOS FP data mentions that the weather forecasts using the GEOS system must be experimental and is for research purpose only.

The main advantages of MERRA-2 and GEOS FP collections over other reanalyses, historical meteorological data, and weather forecasts, can be listed as follows: (i) full compatibility in terms of grid points and meteorological variables, (ii) incorporation of two variables that are very important in wind power prediction – wind speed and direction at the 50 m above ground, (iii) the same data assimilations systems (GEOS), (iv) temporal coverage of MERRA-2 data from 1980 onwards, (v) meteorological forecasts of 10 days ahead, (vi) and free access to the data. Thanks to these advantages, every wind turbine regardless both of the site (either onshore or offshore) and of the starting time of their operation can be used for the study. Moreover, any adaptation of GEOS FP meteorological forecasts to the historical meteorological data (e.g. grid points, altitude, time resolution etc.) collected in MERRA-2 data is unnecessary.

4. Wind turbine output power prediction approach

4.1. Overall framework

A new model based on gradient boosting implementations, historical wind turbine data, MERRA-2 reanalysis and GEOS FP weather forecasts is presented in details. The flowchart of the approach is shown in Figure 4.1. The entire prediction process consists of the following steps: (i) input data collecting, (ii) wind turbine data cleaning, (iii) wind turbine data cleaned and MERRA-2 data integrating, (iv) features engineering, (v) gradient boosting modelling, and (vi) wind turbine output power prediction. The steps (i) – (v) can be executed once in the whole process of prediction or repeated from time to time while the new wind turbine data and MERRA-2 data are accessible. Such updates enable to improve the model performance thanks: (i) the bigger number of data used for model training and (ii) incorporation the loss of wind turbine productivity with time (e.g. ageing, degradation, fatigue). Step (vi) can be repeated daily as new GEOS FP data become available.

Input data are the following: wind turbine dataset, coordinates of wind turbine site, coordinates of MERRA-2 and GEOS FP grid points and, MERRA-2 and GEOS FP datasets. A wind turbine dataset consists of both the measurements (active and reactive power, wind speed and direction measured by anemometer at the top of a nacelle, angle of a nacelle, pitch angle, temperature of turbine components, and many more) and the operational statuses, recorded by SCADA system of the turbine. The average measurements are being recorded in regular periods of time (5 or 10 minutes, depending on the turbine manufacturer). Usually, some of the rows of measurements do not exist because of SCADA errors, turbine outage for safety reasons. In wind turbine output power prediction relying on the methodology presented in the study, only active power is processed. Some of the operational statuses concern the event or state that stops the turbine, e.g. internal faults and failures of components, icing the blades, external grid failures, manual stops, maintenance, calibration the load control and so on. Each status is recorded directly after an event or a state occurred, and is described by its

name, a unique code, and its duration. Time stamps of measurements and operational statues of the turbine in question refer to local times.

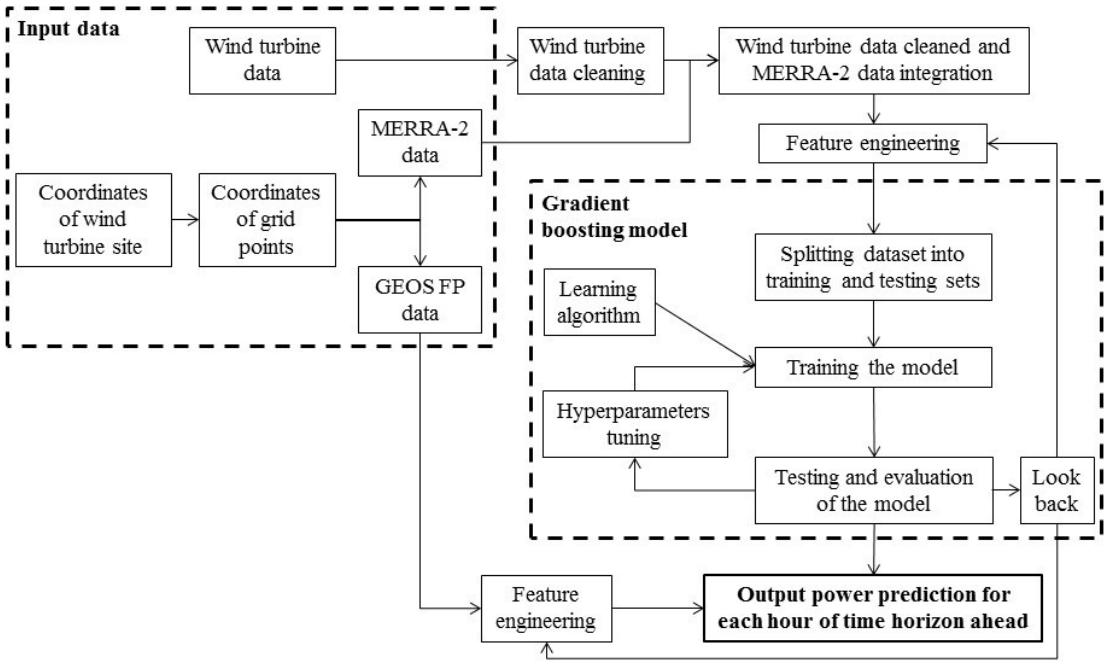

Figure 4.1. Flowchart of the wind turbine output power prediction approach relying on the gradient boosting model

Coordinates of wind turbine site (latitudinal and longitudinal directions) have to be used to determinate the coordinates of MERRA-2 and GEOS FP grid points. The principle of this process is to find the nearest grid points around the wind turbine site coordinates (see Figure 4.2, wind turbine site (a) and grid points: 1 – 4).

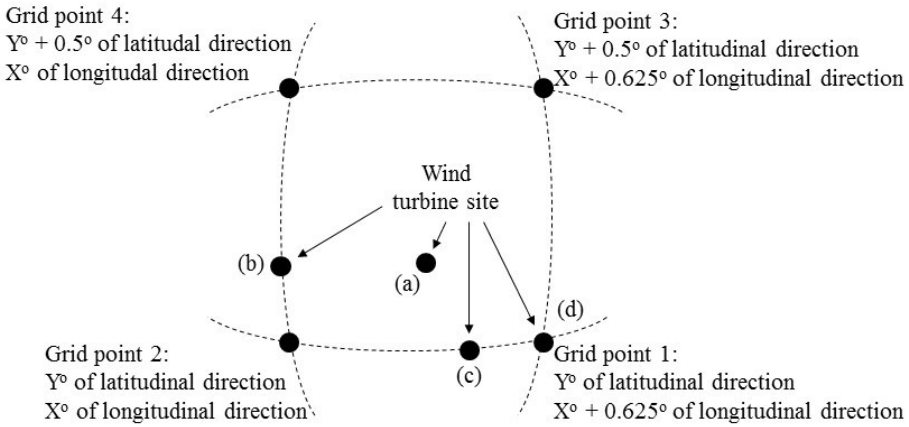

Figure 4.2. The layout of grid points of MERRA-2 and GEOS FP data, and wind turbine site, where Xº and Yº – coordinates of grid point 2

However, if one coordinate (either longitude or latitude) of a wind turbine site and one coordinate of MERRA-2/GEOS FP are the same, the principle is to choose the two nearest grid points, e.g. grid points 2 and 4 (in case of wind turbine site (b)) or grid points 1 and 2 (in case of wind turbine site (c)). And finally, if two coordinates of both wind turbine site and grid point are the same, it is enough to choose this only grid point, e.g. grid point 1 (in case of wind turbine site (d)).

When the grid points are known one can retrieve the meteorological features collected in 'tavg1_2d_slv_Nx' of MERRA-2 and GEOS FP datasets for each grid point (the list of meteorological features gained and other details addressed both datasets are provided in Section 3). The time span of retrieved MERRA-2 data should match the time span of historical wind turbine data in question. The longer the wind turbine data time span is, the better the output power prediction accuracy of the Wind Turbine (WT) is. Temporal coverage of MERRA-2 data from 1980 onwards guarantees the same time span of MERRA-2 and wind turbine data. The number of retrieved GEOS FP data depends on the output power prediction time horizon needed (with a maximum of 240 h ahead).

The units of some the features collected in MERRA-2 and GEOS FP datasets are converted, i.e. pressure from [Pa] into [hPa], humidity from [kg/kg] into [g/kg], temperature from [K] into [^{o}C]. Then one calculates wind speed [m/s] relying on orthogonal velocity components of the wind, U (the zonal velocity – the component of the horizontal wind towards east) and V (the meridional velocity – the component of the horizontal wind towards north), i.e.

$$WS = \sqrt{U^2 + V^2}, \tag{4.1}$$

and meteorological direction in terms of both the degree of range 0 ... 360^{o} based on components U and V

$$WD_{MET}(deg) = \frac{180}{\pi} \cdot atan2(-U, -V), \tag{4.2}$$

and 16 cardinal (and intercardinal) directions relying on $WD_{MET}(deg)$: 'N' – 348.75^{o}–11.25^{o}, 'NNE' – 11.25^{o}–33.75^{o}, 'NE' – 33.75^{o}–56.25^{o}, 'ENE' – 56.25^{o}–78.75^{o}, 'E' – 78.75^{o}–101.25^{o}, 'ESE' – 101.25^{o}–123.75^{o}, 'SE' – 123.75^{o}–146.25^{o}, 'SSE' – 146.25^{o}–168.75^{o}, 'S' – 168.75^{o}–191.25^{o}, 'SSW' – 191.25^{o}–213.75^{o}, 'SW' – 213.75^{o}–236.25^{o}, 'WSW' – 236.25^{o}–258.75^{o}, 'W' – 258.75^{o}–281.25^{o}, 'WNW' – 281.25^{o}–303.75^{o}, 'NW' – 303.75^{o}–326.25^{o}, 'NNW' – 326.25^{o}–348.75^{o}.

Wind turbine data cleaning involves: (i) filling in with the missing rows in original dataset of measurements, (ii) removing the data from some rows and (iii) calculating hourly averages of the data. Completing missing values relies on adding both the rows and their date/time (regular period of time of 5 or 10 min). The number of missing values between two rows of original dataset ranges from one up to hundreds (or even thousands). Removing the data (except date and time) should be carried out relying on operational statuses recorded by the SCADA system of wind turbine. The principle of this process is to remove one or more rows of original dataset of measurements which coincide with time of occurrence and duration of the statuses which concerns the stopping of the turbine. Moreover, if at least one row (time period of 5 or 10 min) of time interval of 1 h (from HH:10 till HH+1h:00, where HH is an hour of the day) is missing or removed, the remaining of rows of this interval should be removed as well, to avoid errors while calculating the hourly averages of the data. The last step of data cleaning consists in calculating of the hourly averages of data (from HH:10 till HH+1h:00). If there is not data in the interval (because of missing values in the original dataset or of removed data) the row affecting this interval will be empty (except for date and time).

While the wind turbine data are cleaned, they can be integrated with MERRA-2 data to constitute the feature dataset. MERRA-2 data must be adapted to local time and synchronized in time with wind turbine data. GEOS FP data must be adapted to local time as well.

Features engineering applies to both cleaned integrated wind turbine data and MERRA-2 data, and GEOS FP data in the same way. It should be carried out before feeding a gradient boosting model with data, and turbine output power prediction, respectively. The main goal is to format input dataset to fit the data-assumption of the model at hand. It can significantly improve the performances of the model. Features engineering consists in: (i) updating the datasets with the columns of the features resulting from given 'Look back' parameter, (ii) categorical features encoding, (iii) data standardization and normalization, and (iv) cyclic features transformation. The details on features engineering are provided in Subsection 5.2.

ML modelling is the process that consists of a few steps, i.e.: (i) splitting the input dataset of the features into training and testing sets, (ii) training the model based on learning algorithm, (iii) hyper-parameters tuning, and (iv) testing and evaluation of the model. The whole dataset should be divided into two subsets – first to train the model and the second to test it. Let us assume that the proportions of the subsets are respectively 80%/20%. Training the model is being carried out relying on: learning algorithm and the training subset of data. As a learning algorithm, three implementations of gradient boosting, i.e. CatBoost, LightBoost and XGBoost have been used. Some details concerning the implementations can be found in Section 3. Before starting the training process, some of the parameters should be set by the user. They are called hyper-parameters and are essential to the performance of the model, the speed and the quality of the learning process. Usually the learning algorithms come with these parameters set to default values that are not necessarily fit for specific problem at hand. In practice, one needs to use appropriate strategies in order to find the best values with respect to the performance criteria. Grid-search is one of such strategies and mostly consists of an exhaustive search on specific parameters values. Having completed the training process, the testing and the evaluation of the model can be carried out. The testing process involves calculating output power based on a trained model and testing subset of features dataset. The evaluation of the model relies on measuring its performances, i.e. statistical relationship between time series of real values of wind turbine output power and time series of the values output power obtained based on the model and testing subset of input data. Two common errors metrics have been used, i.e. RMSE and MAE (see Subsection 5.4). Moreover, the error metrics can be used to correct the hyper-parameters and the 'Look back' parameter, to improve the model performances as much as possible. Having corrected the hyper-parameters and 'Look back' parameter the process of training and testing should be repeated. The gradient boosting model that assures the best performance can be approved and taken for wind turbine output power prediction.

Finally, the resulting model updated with GEOS FP meteorological forecasts for assumed time horizon of prediction returns the time series (point values) of wind turbine output power. Preprocessing the GEOS FP data is the same as MERRA-2 data in terms of features engineering (see Subsection 4.2). The first hour of the time horizon of wind turbine output power prediction and the time horizon depend on: the local time shift (in relation to UTC) and the 'Look back' parameter. For example, if the local time shift is UTC+1 h for winter time and UTC+2 h for summer time (as for CET) and if 'Look back' = 1, then the first hour of power prediction is 02:30 AM CET (winter time) and 03:30 AM CET (summer time), and the time horizon is 239 h ahead. With 'Look back' = 5, the first hour of power prediction is 06:30 AM CET (winter time) and 07:30 AM CET (summer time), and the time horizon is 235 h ahead.

4.2. Feature engineering

The features dataset is expanded with some more columns of data, i.e. lagged features of MERRA-2 data and additionally generated data, and GEOS FP data. Lagged features consist of data from the past (at time intervals t-1, t-2, …, t-n). The number of time intervals

is called ‘Look back’ and must be set by the user, i.e. ‘Look back’ = *n*. The default value of ‘Look back’ may be 1, 2 or 3. Lagged features insertion is motivated by the strong temporal correlation among turbine output power within time interval *t* and meteorological features (e.g. wind speed and direction) in intervals *t*-1, *t*-2, and so on. The criterion to select the value of ‘Look back’ is the best performance of the model that can be achieved within its testing and evaluation.

On the basis of the date-time feature included in the features dataset in the YYYYMMDD:HH format, three other features have been generated, i.e.: hour of the day (0, 1, …, 23), month (1, 2, …, 12), and weather season (winter – from YYYY.12.21:00 till YYYY.03.20:00, spring – from YYYY.03.20:00 till YYYY.06.20:00, summer – from YYYY.06.20:00 till YYYY.09.22:00, and autumn – from YYYY.09.22:00 till YYYY.12.21:00) for each row of features dataset.

Categorical features take the finite set of values as opposite to continuous features. Most gradient boosting implementations, with a few exceptions such as CatBoost, do not support categorical features natively. Thus, they must be encoded before feeding a gradient boosting model with the data at hand. The list of categorical features of our dataset consists of cardinal (intercardinal) wind directions for each grid point in question and time intervals – *t*, *t*-1, *t*-2, …, *t*-*n*. The wind directions are mapped into numbers as follows: ‘N’: 0, ‘NNE’: 1, ‘NE’: 2, ‘NEE’: 3, ‘E’: 4, ‘SEE’: 5, ‘SE’: 6, ‘SSE’: 7, ‘S’: 8, ‘SSW’: 9, ‘SW’: 10, ‘SWW’: 11, ‘W’: 12, ‘NWW’: 13, ‘NW’: 14, ‘NNW’: 15. Data normalization consists of re-scaling the original data provided, so that all values of the features, are within the range 0 – 1. It has been carried out based on the following formula

$$x_{NEW} = \frac{x - x_{MIN}}{x_{MAX} - x_{MIN}}, \tag{4.3}$$

where x_{NEW} – new normalized value, x, x_{MIN} and x_{MAX} – actual, min and max value respectively.

Data normalization is necessary for many machine learning algorithms and especially for LSTM algorithms which are based on neural networks. There are different possibilities to normalize data. There are also many different solutions to normalize or scale the data. Standardization is defined by

$$x_{NEW} = \frac{x - \mu}{\sigma}, \tag{4.4}$$

where x_{NEW} – new normalized value, μ and σ – mean and standard deviation of the original data respectively.

Classically, the min-max normalization (or rescaling) between [-1,1] is defined by

$$x_{NEW} = -1 + \frac{2(x - min\ (x))}{max(x) - min\ (x)}, \tag{4.5}$$

where min(x) and max(x) are respectively the minimum and the maximum value of x.

Both versions were tested and the tests show that the min-max normalization provides the best predictions.

Cyclic features transformation addresses those features whose values are cyclical, i.e. hours of day, days of week, months of year, and wind direction. Their optimal processing can lead to the optimal learning process of the model, i.e. these new features carry the correct information we want the model to learn. Two equations based on trigonometric functions (sin, cos), enable one to create two features corresponding to original feature at hand, i.e.

$$x_{\sin} = \sin\left(\frac{2 \cdot \pi \cdot x}{max\ (x)}\right),\ x_{\cos} = \cos\left(\frac{2 \cdot \pi \cdot x}{max\ (x)}\right), \tag{4.6}$$

where x – the values of original feature.

4.3. Output power prediction accuracy measures

To evaluate the accuracy of a wind turbine output power prediction the following choices have been made: (i) two primary metrics commonly used in the field, i.e. root mean squared error (RMSE) and mean absolute error (MAE), (ii) their relative metrics, i.e. root relative squared error (RRSE) and relative absolute error (RAE) respectively, and (iii) three normalized metrics, i.e. normalized root mean squared error – normalized by the mean of actual data (NRMSE_m), and normalized by the standard deviation of the actual data (NRMSE_sd). RMSE and MAE were used for evaluation of prediction models performance. The measures mentioned above are defined as follows [24]:

- root mean square error, $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|X_i - \hat{X}_i\right|^2}$,
- relative root squared error, $RRSE = \sqrt{\frac{\sum_{i=1}^{N}\left|X_i - \hat{X}_i\right|^2}{\sum_{i=1}^{N}\left|X_i - \overline{X}\right|^2}}$,
- normalized root mean squared error (normalized by the standard deviation of the actual data), $NRMSE_{sd} = \frac{RMSE}{sd}$,
- normalized root mean squared error (normalized by mean of actual data), $NRMSE_m = \frac{RMSE}{\overline{X}}$,
- mean absolute error, $MAE = \frac{1}{N}\sum_{i=1}^{N}\left|X_i - \hat{X}_i\right|$,
- relative absolute error, $RAE = \frac{\sum_{i=1}^{N}\left|X_i - \hat{X}_i\right|}{\sum_{i=1}^{N}\left|X_i - \overline{X}\right|}$,

where: X_i – ith actual value, $\hat{X}_i$ – ith value obtained while testing the model or output power predicting, $\overline{X}$ – mean of the actual values, X – the set of actual values, sd – standard deviation of actual values, N – size of the dataset.

RRSE and NRMSE_sd are the same in terms of mathematical formula. The smaller the values of the metrics, the more efficient gradient boosting models and output power prediction results are. Let us assume the criterion of acceptance of output power prediction is both relative and normalized errors do not excide 1. According to the performance requirement in the functional specification of wind power forecasting system reported by

State Grid Corporation of China, article IV, the RMSE of short-term forecasting of a single wind farm should be less than 20%. The RMSE of the predicted value of 4 h should be less than 15% [4].

5. Case study

5.1. Wind turbine operation data

The case study addresses a prediction of output power of a wind turbine E-53 by ENERCON. The rated power of the turbine is 800 kW. It is located in the North-Eastern of Poland. Original wind turbine dataset consists of the measurements of turbine parameters and its operational statuses, both recorded in SCADA system over period 2014.09.01–2020.11.30. Resolution of the data is 10 min (interval averages). Data cleaning of original dataset of measurements has involved: filling in with the missing rows, removing the data in some rows (data of 10 min resolution) based on operational statuses of the turbine, and calculating hourly averages of the data (see the rules in Subsection 4.1). The total number of hourly averaged measurement rows is 54787. The hourly averages of the data have been calculated to be compatible with MERRA-2 data and finally – to constitute the features dataset. For comparison with the dataset of hourly averages calculated relying on the rule described in Subsection 4.1, a second dataset was created in which hourly averaged values are obtained without applying this rule. The number of missing records, mean, standard deviation, variance of wind turbine output power and correlation between output power and wind speed measured by anemometer of the turbine, in both datasets are provided in Table 5.1.

Table 5.1. Parameters that summarize the wind turbine output power data, without and with applying the rule of hourly averaged data calculation

Parameter	Without applying the rule	With applying the rule
number of missing values (percentage of total number of data)	1150 (2.1%)	3565 (6.5%)
mean [kW]	178.92	182.61
standard deviation [kW]	180.83	180.63
variance	32698.8	32626.5
coefficient of Pearson correlation between output power and wind speed	0.914	0.944

As provided in Table 5.1 the number of missing values in both datasets is not large respectively – 2.1% and 6.5% of the total number of data. The application of the rule results in a threefold increase of missing data. Since the total number of records is huge, one can omit the missing values in processing with output power prediction without any significant consequences. Following this rule resulted in: (i) both mean of output power and coefficient of correlation between output power and wind speed increase and (ii) both a standard deviation and a decrease in the variance.

Figure 5.1 presents the scatter plots and histograms of two features: wind turbine output power and wind speed measured by turbine's anemometer. The main difference between both scatter plots is the number of data points below the common sigmoidal shaped power curve of the turbine – mostly in or near the horizontal axis where the output power is 0. The histograms of output power (and wind speed) are very similar to each other in terms of bins distribution. The only difference is the frequency of data in the bins. It can be noted that the distribution of wind speed tends to be a normal distribution and the mean is much lower (~ 5 - 6 m/s) than the rated wind speed (following manufacturer data – 12 m/s), for both datasets. This implies that a wind turbine spends the majority of its operating time much below the rated power (800 kW).

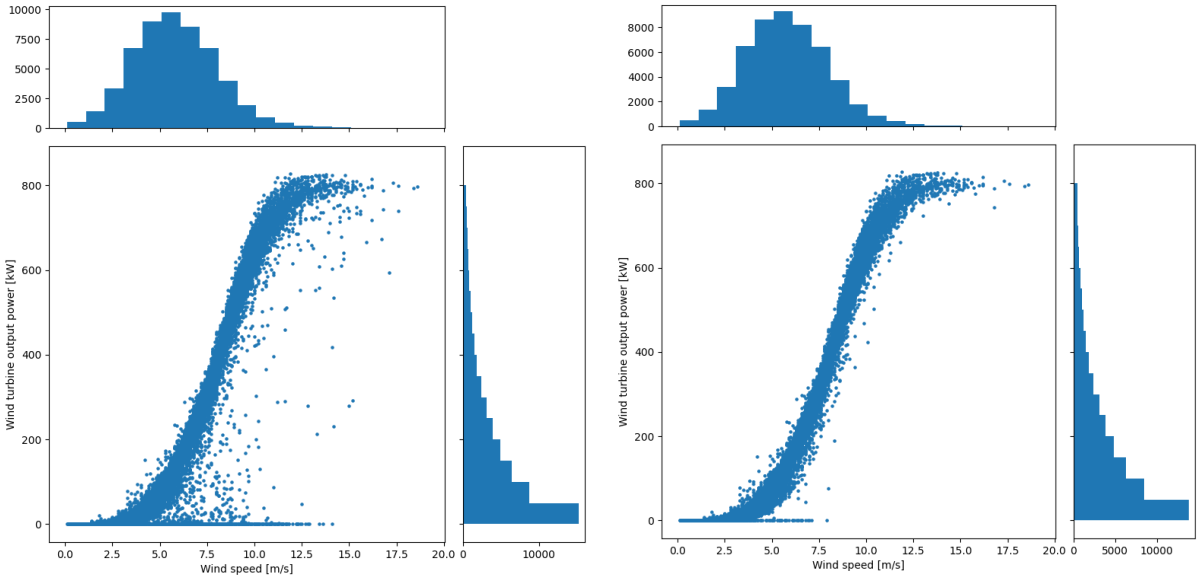

Figure 5.1. Scatterplots and histograms of wind turbine output power and wind speed measured by turbine's anemometer within the period 2014.09.01-2020.11.30, without (left) and with (right) applying the rule of hourly averaged data calculation

Following these experiments, one can conclude that the performance of the models and the predictions accuracy are essentially better for the dataset that was created using the rule of hourly averaged data calculation. Only the prediction results based on a dataset created with applying the rule of hourly averaging data are presented in the further part of the work.

5.2. MERRA-2 and GEOS FP data

Coordinates of the wind turbine site (latitudinal and longitudinal directions) were used to determinate the coordinates of MERRA-2 and GEOS FP grid points based on the rule presented in Subsection 5.1. The turbine site corresponds to case (a) in Figure 4.2, i.e. the coordinates of the four nearest grid points were found. The distances between wind turbine and grid points of MERRA-2 reanalysis and GEOS FP data are as follows: WT and grid point 1 = 38.31 km, WT and grid point 2 = 30.51 km, WT and grid point 3 = 39.02 km and WT and grid point 4 = 31.5 km. The time span of retrieved MERRA-2 data is the same as the one found in the wind turbine data, i.e. 2014.09.01 – 2020.11.30. Collected meteorological features are listed in Subsection 5.1. MERRA-2 dataset was adapted to local time (CET), i.e. UTC+1 for winter time and UTC+2 h for summer time and finally integrated with cleaned wind turbine data. GEOS FP datasets were retrieved for many days during the following period 15.08.2020 – 15.02.2021, but it was decided to proceed with the forecasts for 14 randomly selected days, i.e.: 28.04.2020, 31.08.2020, 07.09.2020, 23.09.2020, 05.10.2020, 19.10.2020, 03.11.2020, 27.11.2020, 09.12.2020, 28.12.2020, 04.01.2021, 25.01.2021, 02.02.2021, 08.02.2021. The only criterion of the day selection was to take two days of one month, in a proportional way among different seasons. GEOS FP datasets were adopted to local time (CET) as well. Each GEOS FP dataset consists of 240 hourly averaged meteorological features starting from the announced above.

5.3. Features engineering and ML modelling

The integrated dataset of cleaned wind turbine data and MERRA-2 data were upgraded with the columns of meteorological features given a 'Look back' parameter. One started with 'Look back' = 1, thus lagged features consist of the data shifted by 1 h back (*t*-1). Moreover, three other features were generated and mapped into numbers and incorporated into the

dataset, i.e. hour of the day, month of the year, and season. Cardinal and intercardinal wind directions (that constitute categorical feature) were mapped into the numbers as well. All the features were normalized and standardized relying on formula (1) and (2), respectively. The cyclic features (hours of the day, days of the month, month of the year, and wind direction) were transformed based on formula (3). GEOS FP datasets were processed in the same way as the integrated dataset. The one proceeded with different 'Look back' parameter values, i.e. from 1 up to 7. After changing the 'Look back' both the integrated dataset and GEOS FP datasets were reconfigured in order to take into account the new lagged features. All meteorological features retrieved from MERRA-2 and GEOS FP relating to a given grid point were indexed with that point number, e.g. WS50M1 concerns wind speed at 50 meters at a point 1.

Three GB implementations were used for training the model and wind turbine output power prediction, i.e. CatBoost, LightBoost, and XGBoost (see Section 2). Moreover, in order to compare the performance of GB models and their predicting accuracy with other ML algorithms it was decided to introduce three other models that rely on: LSTM, Decision Tree and Random Forest (see Section 2). The basic criterion for selecting these approaches was the need to ensure the same idea of prediction process in terms of: learning from the data, input dataset that consists of wind turbine data and MERRA-2 data for training the model, and GEOS FP weather forecast data for prediction of the wind turbine output power.

All the models in question were trained using all features included into the input dataset. Hyper-parameters optimization and tuning were performed for each learning algorithm. Concerning GB models the grid-search approach was used for hyper-parameters optimization and tuning to ensure their optimal performance. The models' performance was measured thanks to two error metrics, i.e. RMSE and MAE. All the models were trained and tested using identical training and testing datasets. The same GEOS FP meteorological forecast datasets were used for wind turbine output power prediction relying on all the models (see Subsection 5.2).

5.4. Results and discussion

Following many experiments during the models' training (different 'Look back' parameter assumption, hyper-parameters optimization and tuning, different architectures of LSTM model) and their testing and evaluation, the best performance of the models in terms of both RMSE and MAE is achieved for 'Look back' = 5 and summarized in Table 5.2. Since 'Look back' = 5 the total number of processed features is 501.

Table 5.2. The learning algorithms performance (the best score is in bold)

Learning algorithm	RMSE [kW]	MAE [kW]
CatBoost	**76.18**	**54.87**
LightBoost	76.84	55.24
XGBoost	77.02	55.61
LSTM	78.73	57.85
DecisionTree	111.26	79.38
RandomForest	77.97	56.14

Table 5.2 shows that GB models have the best performances as compared to the other models. Among three GB models, their performance is comparable but CatBoost implementation is at the top (best results of RMSE and MAE in bold). On the other hand, the Decision Tree model has the worst performance. In regard to the LSTM model and to the Random Forest model, their performance is very close to that of the GB models.

The CatBoost model assures the best performance (see Table 5.21). With this model, one can see in Figure 5.2 the 10 best features. These 10 features all concern wind speeds. Four of them refer to an altitude of 50 m above ground at all grid points (WS50M2, WS50M4, WS50M3, WS50M1). Three wind speeds concern an altitude of 10 m above zero plane displacement high at grid points 3 (WS10M3) and 2 (WS10M2), and 4 (WS10M4). Moreover there are three lagged features, i.e. two for *t*-1 (WS50M4_T1 and WS50M2_T1), and one for *t*-2 (WS50M4_T2). Among these features, there are three that dominate in most – WS50M2, WS50 M4 and WS50M4_T1.

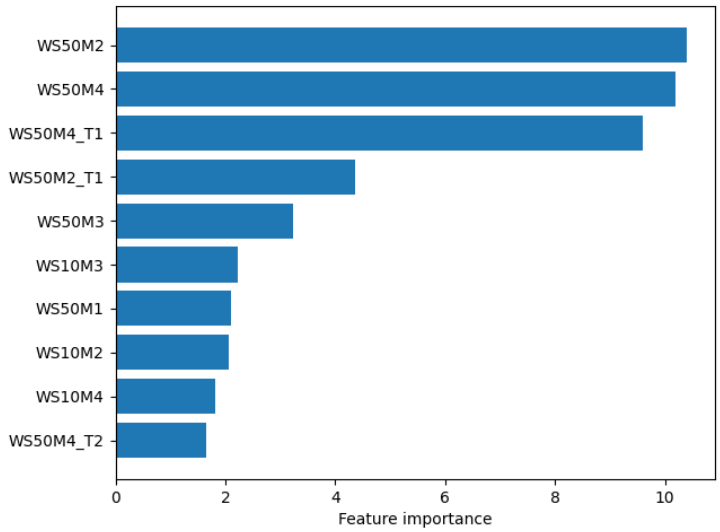

Figure 5.2. Features importance ranking for CatBoost model

The correlation among wind turbine output power and ten meteorological features that extremely important for the CatBoost model are shown in Figure 5.3.

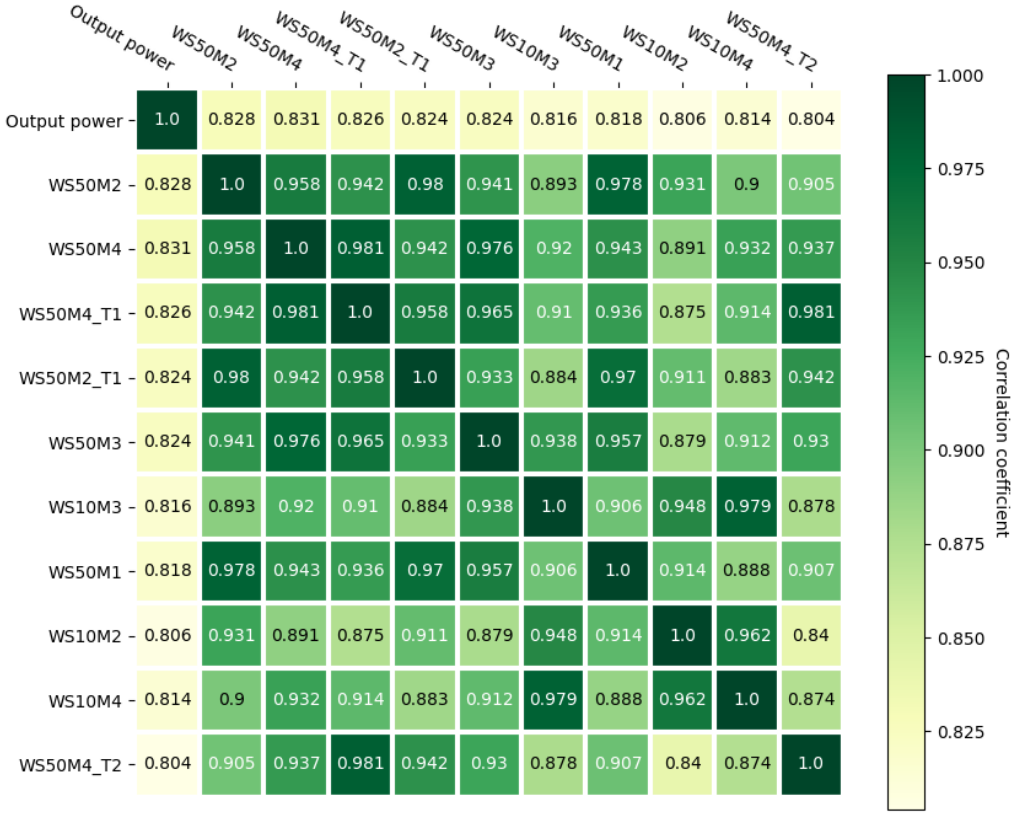

Figure 5.3. Heatmap of the features for CatBoost model

Figure 5.3 confirms a strong correlation: (i) between wind turbine output power and relevant meteorological features (coefficient of Pearson correlation from 0.806 for WS10M2 up to 0.831 for WS50M4), and (ii) among all of meteorological features (from 0.84 between WS10M2 and WS50M4_T2 up to 0.981 between WS50M4_T1 and WS50M4_T2).

Following the investigation of different number of features used in models training process carried out by the authors, it can be concluded that the best performance of the models is assured by the use of all the available features instead of a few ones of higher importance in the ranking.

Since 'Look back' = 5 assures the best performance of the models in question, the time horizon of the wind turbine output power prediction is 235 h ahead. The error measures of wind turbine output power prediction calculated for all 14 days (i.e. 28.04.2020, 31.08.2020, 07.09.2020, 23.09.2020, 05.10.2020, 19.10.2020, 03.11.2020, 27.11.2020, 09.12.2020, 28.12.2020, 04.01.2021, 25.01.2021, 02.02.2021, 08.02.2021) relying on 6 models are provided in Appendix A, B, and C, for time horizon prediction of 48, 120 and 235 h ahead, respectively. The appendixes contain the calculated error measures as follows: RMSE, RRSE (NRMSE_sd), NRMSE_m, MAE, and RAE. The best scores of the errors are in bold. Following the appendixes the range of error values calculated for individual predictions is very large. The most spectacular comparison of the range of error values refers to the relative errors, i.e. RRSE, NRMSE_m and RAE. For example, RRSE ranges from about 0.3 (for 08.02.2021) to greater than 1.4 (for 09.12.2020). The longer the prediction time horizon is, the smaller the error range is, and the higher the lower limit is contrary to what happens in the case of short-term predictions.

Table 5.3 shows the number of accepted predictions relying on the models in question assuming that the acceptance criteria of wind turbine output power prediction are consistent with those given in Subsection 5.4 (RMSE and MAE lower than 20% of nominal power of wind turbine, and RRSE(NRMSE_sd), NRMSE_m, NRMSE_mm and RAE lower than 1).

Table 5.3. The number of accepted predictions relying on all the models considered (the best scores are in bold)

Learning algorithm	Number of accepted predictions for a time horizon (% of total number of predictions):		
	48 h	120 h	235 h
CatBoost	**11 (78.57%)**	9 (64.29%)	**6 (42.86%)**
XGBoost	10 (71.43%)	**10 (71.43%)**	**6 (42.86%)**
LightBoost	10 (71.43%)	9 (64.29%)	**6 (42.86%)**
LSTM	8 (57.14%)	8 (57.14%)	5 (35.71%)
Decision Tree	5 (35.71%)	4 (28.57%)	2 (14.29%)
Random Forest	**11 (78.57%)**	**10 (71.43%)**	5 (35.71%)

In Table 5.3 the best numbers of accepted predictions are in bold. With this table one can conclude as follows:

- the CatBoost model and the Random Forest model provide the most acceptable predictions for 48 h ahead (78.57%), XGBoost model and Random Forest – for 120 h ahead (71.43%), and GB models – for 235 h ahead (42.86%),
- there are not any learning algorithms that produce 100% of accepted predictions,
- the longer the time horizon of prediction is, the lower the number of accepted predictions is.

A comparison of the wind turbine output power prediction relying on all the models is shown in Table 5.4, giving time horizons of 48 h, 120 h, and 235 h. The values of absolute

errors (in kW and % of wind turbine rated power) are averaged over fourteen absolute errors presented in Appendix A, B and C.

Table. 5.4. Averaged absolute errors (RMSE and MAE) of the wind turbine output power prediction relying on all the models, given time horizon of 48 h, 120 h, and 235 h (the best scores are in bold)

Error measure	Learning algorithm	48 h ahead		120 h ahead		235 h ahead	
		Average value	% of rated power	Average value	% of rated power	Average value	% of rated power
RMSE	CatBoost	**104.03**	**13.00**	**113.30**	**14.16**	169.05	21.13
	XGBoost	104.77	13.10	114.85	14.36	168.66	21.08
	LightBoost	105.28	13.16	114.85	14.36	169.47	21.18
	LSTM	116.98	14.62	126.14	15.77	168.36	21.04
	DecisionTree	135.96	17.00	143.88	17.98	186.48	23.31
	RandomForest	106.53	13.32	114.47	14.31	**165.96**	**20.75**
MAE	CatBoost	80.40	10.05	**82.79**	**10.35**	117.34	14.67
	XGBoost	82.10	10.26	84.48	10.56	118.08	14.76
	LightBoost	**79.45**	**9.93**	83.14	10.39	116.89	14.61
	LSTM	92.94	11.62	94.49	11.81	120.02	15.00
	DecisionTree	102.24	12.78	102.21	12.78	130.60	16.33
	RandomForest	82.82	10.35	85.36	10.67	**116.17**	**14.52**

In Table 5.4 the best scores of the averaged errors are in bold. It shows that the GB models provide the best predicting accuracy for a time horizon of 48 h and 120 h ahead. The differences in error values are slight in respect to three GB implementations. The best prediction results are provided by the CatBoost model (the best scores of: (i) RMSE for 48 h and 120 h, and (ii) MAE for 120 h). The LightBoost model can be considered the second ranked model (the best score of MAE for 48 h). The accuracy of the Random Forest model is similar to the GB models but slightly less efficient as far as 48 h and 120 h predictions are concerned. The Random Forest model turned out to be the best for predictions of 235 h ahead (the best scores of RMSE and MAE). Its accuracy is only slightly better than that of the GB models and the LSTM model. The Decision Tree model seems to be the worst one regardless of the time horizon of prediction.

There are two reasons for an imperfect output power prediction, i.e. the performance of the learning algorithm (see the errors in Table 5.2) and the differences between GEOS FP meteorological forecasts against MERRA-2 data.

Since the CatBoost model assures the best performance (see Table 5.2) and demonstrates one of the best accuracy of the wind turbine output power prediction, Figure 5.4 depicts the plots of true and predicted output power for 10 selected days.

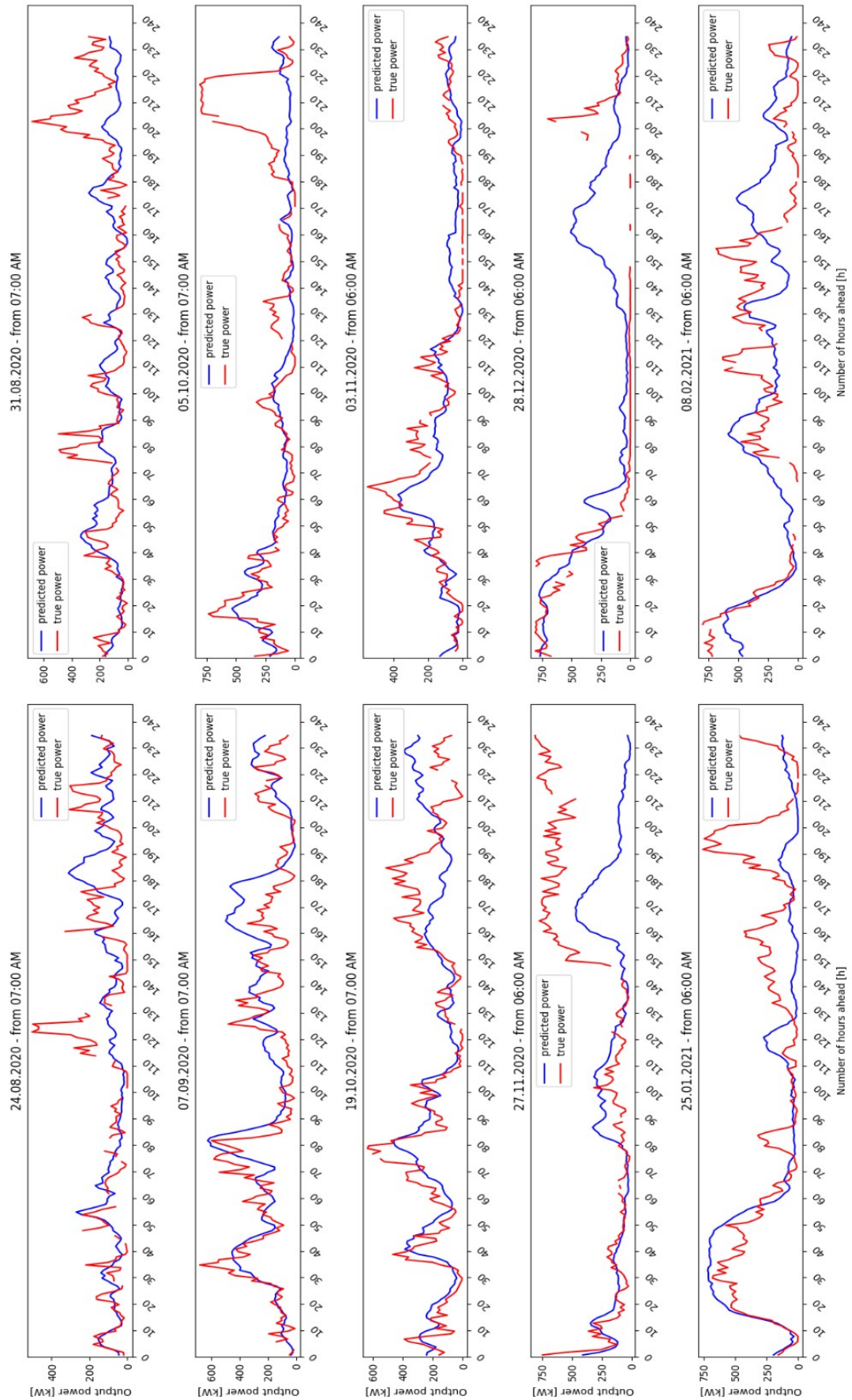

Figure 5.4. Comparison of true and predicted wind turbine output power for 235 h ahead (the day and time of prediction are provided above the plot) obtained relying on CatBoost model

6. Conclusions

This work presents an in-depth investigation of the wind turbine output power prediction by using historical data (both wind turbine data and MERRA-2 data), GEOS FP meteorological forecast and GB algorithm. Three implementations of the GB algorithm were compared to create optimized predictive models, i.e. CatBoost, XGBoost, and LightBoost. In order to measure the performances of the GB models and the accuracy of the predictions other ML algorithms were tested: LSTM, Decision Tree and Random Forest. The basic criterion to select these approaches was the need to ensure the same idea of prediction process in terms of: learning from the data, input dataset that consists of wind turbine data and MERRA-2 data for training the model, and GEOS FP weather forecast data for prediction of the wind turbine output power. The approach of the wind turbine output power prediction was achieved in several steps defined by: input data preprocessing and cleaning, feature engineering, training and testing the models and finally obtaining the output power prediction for each hour of time horizon ahead. To maximize the performance of the models, hyper-parameter tuning was used by manual and grid search (grid search optimization process was applied to GB models). The approach presented in the work is applicable regardless of the site (onshore, offshore) of wind turbine for which the prediction is made. The type and the technical specification of the investigated turbine are also irrelevant. The only thing that may have an impact on the accuracy of the prediction is to ensure that the models training is repeated as new data about both the wind turbine operation and MERRA-2 data is obtained and incorporated into input dataset. To sum up, the following facts can be highlighted based on the results of the wind turbine output power prediction:

- data must be preprocessed and cleaned before they become input data in the model training process,
- the greater number of features in input dataset is, the better the performance of the models is; the inclusion of a few additional categorical features (season, hour of the day, cardinal/intercardinal wind direction) in the input dataset can also improve the performance of the models,
- the GB models achieve the best performance compared to the other models investigated,
- assuming the criteria of power prediction acceptance, the GB models demonstrate the evidently higher number of accepted predictions as compared to the LSTM model and the Decision Tree model, regardless of the time horizon of the prediction; the number of accepted predictions relying on the Random Forest model is comparable to the GB models,
- The GB models (especially the CatBoost implementation) provide the smallest averaged absolute errors of the wind turbine output power prediction for a time horizon of 48 h and 120 h ahead,
- The GB models and the Random Forest model provide the comparable averaged absolute errors of the wind turbine output power prediction for 235 h ahead.

The output power prediction errors obtained relying on the approach and learning algorithms (especially GB models) presented in this work for the assumed time horizons (48 h, 120 h and 235 h) ahead are at a much lower level than prediction errors derived thanks to models proposed by other authors.

Acknowledgement

The MERRA-2 and GEOS data used in this study have been provided by the Global Modeling and Assimilation Office (GMAO) at NASA Goddard Space Flight Center through the online data portal in the NASA Center for Climate Simulation. The authors thank the owner of the

wind turbine E-53 for providing wind turbine data. This work has also partially been funded by the EIPHI Graduate School (contract "ANR-17-EURE-0002").

Appendix A. Wind turbine output power prediction errors for prediction horizon of 48 h (the best scores are in bold)

Error measure	Learning algorithm	Day of prediction													
		24.08.2020	31.08.2020	07.09.2020	23.09.2020	05.10.2020	19.10.2020	03.11.2020	27.11.2020	09.12.2020	28.12.2020	04.01.2021	25.01.2021	02.02.2021	08.02.2021
RMSE [kW]	CatBoost	47.44	57.83	74.4	84.63	**107.09**	92.45	40.36	89.31	116.13	116	155.41	142.44	207.19	125.78
	XGBoost	49.77	59.37	73.37	97.18	110.66	91.55	41.14	**86.58**	117.19	**107.13**	177.6	151.9	**199.99**	103.29
	LightBoost	**46.5**	67.43	74.2	96.84	112.21	85.96	40.97	90.82	**113.22**	113.47	168.16	137.81	214.35	112.03
	LSTM	55.64	65.31	102.63	**76.81**	111.29	101.21	**39.89**	96.12	221.32	129.56	177.92	173.35	200.8	**85.8**
	DecisionTree	73.05	75.74	108.87	141.22	180.46	**84.57**	79.13	106.93	134.27	172.95	162.52	173.59	257.3	152.9
	RandomForest	54.05	**56.22**	**68.7**	96.94	123.48	91.64	43.44	100.89	122	124.22	**131.71**	**117.07**	243.49	117.54
RRSE (NRMSE_sd)	CatBoost	0.814	0.67	0.444	1.072	**0.64**	0.786	0.605	0.699	1.466	0.831	0.773	0.686	0.764	0.429
	XGBoost	0.854	0.687	0.438	1.231	0.661	0.779	0.616	**0.677**	1.479	**0.767**	0.883	0.731	**0.738**	0.352
	LightBoost	**0.798**	0.781	0.443	1.226	0.67	0.731	0.614	0.71	**1.429**	0.812	0.836	0.663	0.791	0.382
	LSTM	0.955	0.756	0.613	**0.973**	0.665	0.861	**0.598**	0.752	2.794	0.928	0.885	0.834	0.741	**0.292**
	DecisionTree	1.254	0.877	0.65	1.789	1.078	**0.719**	1.185	0.837	1.695	1.238	0.808	0.836	0.949	0.521
	RandomForest	0.928	**0.651**	**0.41**	1.228	0.738	0.779	0.651	0.789	1.54	0.889	**0.655**	**0.563**	0.898	0.401
NRMSE_m	CatBoost	0.63	0.476	0.333	0.554	**0.346**	0.546	0.469	0.534	0.457	0.177	0.774	0.369	0.758	0.335
	XGBoost	0.661	0.489	0.328	0.637	0.358	0.541	0.478	**0.518**	0.461	**0.163**	0.884	0.394	**0.732**	0.275
	LightBoost	**0.618**	0.555	0.332	0.634	0.363	0.508	0.476	0.544	**0.445**	0.173	0.837	0.357	0.784	0.298
	LSTM	0.739	0.538	0.459	**0.503**	0.36	0.598	**0.463**	0.575	0.87	0.197	0.886	0.45	0.735	**0.229**
	DecisionTree	0.971	0.623	0.487	0.925	0.583	**0.499**	0.919	0.64	0.528	0.263	0.809	0.45	0.942	0.407
	RandomForest	0.718	**0.463**	**0.308**	0.635	0.399	0.541	0.504	0.604	0.48	0.189	**0.656**	**0.304**	0.891	0.313
MAE [kW]	CatBoost	34.17	**42.01**	52.63	68.31	90.47	71.67	32.43	61.08	92.22	89.11	142.88	111.45	141.94	95.27
	XGBoost	37.12	43.32	53.88	79.19	94.35	72.53	33.05	**59.97**	94.09	**80.74**	165.5	119.17	140.1	76.44
	LightBoost	**33**	47.8	**50.46**	75.91	91.22	65.74	**31.88**	61.55	**87.87**	84.89	149.56	105.91	147.95	78.58
	LSTM	46.74	48.78	66.05	**64.1**	**88.37**	85.75	32.31	68.74	196.78	93.12	167.52	139.38	**138.77**	**64.71**
	DecisionTree	60.4	55.53	82.32	103.27	144.54	**64.95**	61.23	76.01	105.44	141.3	130.34	138.84	161.15	106.09
	RandomForest	42.97	43.07	51.94	79.81	101.63	74.62	36.45	64.62	102.56	100.38	**113.5**	**95.79**	168.42	83.7
RAE	CatBoost	0.741	**0.569**	0.371	1.076	0.691	0.768	0.585	0.729	1.447	0.756	0.819	0.625	0.579	0.357
	XGBoost	0.804	0.586	0.38	1.247	0.721	0.777	0.596	**0.715**	1.477	**0.685**	0.949	0.669	0.571	0.286
	LightBoost	**0.715**	0.647	**0.356**	1.195	0.697	0.704	**0.575**	0.734	**1.379**	0.721	0.857	0.594	0.603	0.294
	LSTM	1.013	0.66	0.466	**1.009**	**0.675**	0.919	0.583	0.82	3.088	0.791	0.96	0.782	**0.566**	**0.242**
	DecisionTree	1.309	0.752	0.581	1.626	1.105	**0.696**	1.105	0.907	1.655	1.199	0.747	0.779	0.657	0.397
	RandomForest	0.931	0.583	0.366	1.257	0.777	0.8	0.658	0.771	1.61	0.852	**0.651**	**0.537**	0.687	0.314

Appendix B. Wind turbine output power prediction errors for prediction horizon of 120 h (the best scores are in bold)

Error measure	Learning algorithm	Day of prediction													
		24.08.2020	31.08.2020	07.09.2020	23.09.2020	05.10.2020	19.10.2020	03.11.2020	27.11.2020	09.12.2020	28.12.2020	04.01.2021	25.01.2021	02.02.2021	08.02.2021
RMSE [kW]	CatBoost	**54**	**99.73**	113.64	197.57	**84.09**	97.13	**77.44**	94.99	94.64	106.56	97.9	125.17	164.24	179.14
	XGBoost	57.28	100.87	114.14	205.23	88.07	94.09	77.97	98.65	97.99	**102.86**	112.2	132.16	153.78	172.63
	LightBoost	57.13	101.99	110.61	210.96	85.74	**93.35**	89.25	**94.17**	**92.67**	104.97	105.53	119.92	167.11	174.52
	LSTM	57.5	110.71	118.49	199	93.23	104.5	79.85	112.73	173.85	118.65	111.68	138.63	**153.65**	193.43
	DecisionTree	85.26	125.92	144.16	208.73	136.57	127.94	127.46	129.48	111	150.01	103.46	161.85	211.35	191.11
	RandomForest	57.32	102.05	**108.61**	**192.94**	97.26	98.27	85.86	95.71	99.42	110.68	**86.13**	**107.6**	191.61	**169.07**
RRSE (NRMSE_sd)	CatBoost	**0.859**	**0.946**	0.683	1.172	**0.52**	0.661	**0.609**	0.972	0.725	0.33	0.686	0.602	0.743	0.752
	XGBoost	0.911	0.957	0.686	1.218	0.545	0.64	0.613	1.01	0.751	**0.318**	0.786	0.636	0.696	0.725
	LightBoost	0.909	0.967	0.665	1.252	0.53	**0.635**	0.701	**0.964**	**0.71**	0.325	0.74	0.577	0.756	0.733
	LSTM	0.915	1.05	0.712	1.181	0.577	0.711	0.627	1.154	1.333	0.367	0.783	0.667	**0.695**	0.813
	DecisionTree	1.357	1.194	0.866	1.238	0.845	0.871	1.002	1.325	0.851	0.464	0.725	0.778	0.956	0.803
	RandomForest	0.912	0.968	**0.653**	**1.145**	0.602	0.669	0.675	0.98	0.762	0.342	**0.604**	**0.518**	0.867	**0.71**
NRMSE_m	CatBoost	**0.664**	**0.759**	0.55	0.694	**0.474**	0.497	**0.439**	0.753	0.495	0.388	1.009	0.542	0.911	0.551
	XGBoost	0.704	0.768	0.553	0.721	0.496	0.481	0.442	0.782	0.513	**0.375**	1.156	0.572	0.853	0.531
	LightBoost	0.702	0.776	0.536	0.741	0.483	**0.478**	0.506	**0.746**	**0.485**	0.383	1.087	0.519	0.926	0.537
	LSTM	0.707	0.843	0.574	0.699	0.525	0.535	0.453	0.894	0.91	0.432	1.151	0.6	**0.852**	0.595
	DecisionTree	0.971	0.959	0.698	0.733	0.77	0.655	0.722	1.026	0.581	0.547	1.066	0.7	1.172	0.588
	RandomForest	0.705	0.777	**0.526**	**0.678**	0.548	0.503	0.487	0.759	0.52	0.403	**0.888**	**0.466**	1.062	**0.52**
MAE [kW]	CatBoost	**39.71**	**70.36**	83.26	**141.55**	67.61	75.61	59.25	70.35	75.89	76.39	65.23	93.2	97.03	143.56
	XGBoost	42.61	72.34	85.64	150.64	72.43	73.87	**59.23**	73.84	78.74	**70.64**	76.02	98.5	**92.41**	135.82
	LightBoost	42.51	72.29	**79.44**	157.8	**66.32**	**70.85**	67.67	70.29	**72.99**	76.29	68.35	88.41	98.05	132.69
	LSTM	45.51	79.05	84.84	146.12	74.66	84.95	60.43	86.21	144.22	89.15	78.94	105.78	92.63	150.34
	DecisionTree	64.22	84.74	104.95	162.1	102.97	90.05	98.02	92.37	77.21	108.85	62.26	116.31	124.32	142.53
	RandomForest	44.88	73.91	83.98	148.59	79.44	77.54	66.38	**68.47**	82.11	87.25	**57.53**	**81.91**	112.84	**130.19**
RAE	CatBoost	**0.792**	**0.861**	0.599	**0.973**	0.56	0.653	0.57	1.119	0.665	0.254	0.64	0.503	0.541	0.731
	XGBoost	0.85	0.885	0.616	1.036	0.6	0.638	**0.569**	1.175	0.689	**0.235**	0.745	0.531	**0.516**	0.691
	LightBoost	0.848	0.884	**0.571**	1.085	**0.549**	**0.612**	0.651	1.118	**0.639**	0.254	0.67	0.477	0.547	0.675
	LSTM	0.908	0.967	0.61	1.005	0.619	0.734	0.581	1.372	1.263	0.296	0.774	0.571	0.517	0.765
	DecisionTree	1.281	1.037	0.755	1.114	0.853	0.778	0.942	1.47	0.676	0.362	0.61	0.627	0.694	0.726
	RandomForest	0.895	0.904	0.604	1.022	0.658	0.67	0.638	**1.089**	0.719	0.29	**0.564**	**0.442**	0.63	**0.663**

Appendix C. Wind turbine output power prediction errors for prediction horizon of 235 h (the best scores are in bold)

Error measure	Learning algorithm	Day of prediction													
		24.08.2020	31.08.2020	07.09.2020	23.09.2020	05.10.2020	19.10.2020	03.11.2020	27.11.2020	09.12.2020	28.12.2020	04.01.2021	25.01.2021	02.02.2021	08.02.2021
RMSE [kW]	CatBoost	**102.71**	129.86	131.57	196.01	220.6	128.43	64.27	320.36	255.5	153.68	108.56	198.56	149.22	207.42
	XGBoost	105.66	128.23	129.73	198.17	220.37	130.29	**63.01**	319.88	249.5	151.26	112.75	197.96	**147.63**	206.79
	LightBoost	106.26	134.28	120.13	199.59	221.35	130	70.09	327.77	249.64	149.34	113.95	193.46	150.68	205.99
	LSTM	102.74	132.33	**118.52**	190.54	**218.12**	127.05	64.54	**284.31**	275.38	168.35	113	194.61	152.71	214.79
	DecisionTree	123.64	147.46	143.26	198.07	225.67	148.2	98.85	344.52	251.14	175.37	121.05	220.93	193.34	219.25
	RandomForest	103.55	**126.71**	121.31	**188.92**	219.8	**125.87**	68.24	328.23	**238.26**	**146.01**	**104.92**	**188.81**	167.65	**195.2**
RRSE (NRMSE_sd)	CatBoost	**1.121**	1.04	0.946	1.359	1.041	0.938	0.544	1.162	1.983	0.545	0.615	0.987	0.602	0.918
	XGBoost	1.167	1.027	0.933	1.374	1.04	0.952	**0.533**	1.16	1.936	0.537	0.639	0.984	**0.596**	0.915
	LightBoost	1.173	1.075	0.864	1.384	1.045	0.95	0.593	1.188	1.938	0.53	0.645	0.962	0.608	0.911
	LSTM	1.125	1.056	**0.849**	1.31	**1.027**	0.923	0.541	**1.049**	2.121	0.587	0.637	0.966	0.611	0.947
	DecisionTree	1.365	1.181	1.03	1.373	1.065	1.082	0.836	1.249	1.949	0.622	0.686	1.098	0.78	0.97
	RandomForest	1.143	**1.015**	0.872	**1.31**	1.038	**0.919**	0.577	1.19	**1.849**	**0.518**	**0.594**	**0.939**	0.677	**0.863**
NRMSE_m	CatBoost	**1.012**	0.896	0.703	0.839	1.121	0.666	0.578	0.994	1.607	0.766	0.876	0.893	0.645	0.815
	XGBoost	1.041	0.884	0.693	0.848	1.12	0.676	**0.567**	0.992	1.569	0.753	0.91	0.891	**0.638**	0.813
	LightBoost	1.047	0.926	0.642	0.854	1.125	0.674	0.63	1.017	1.57	0.744	0.92	0.871	0.651	0.81
	LSTM	1.014	0.923	**0.628**	0.815	**1.092**	0.655	0.582	**0.905**	1.723	0.824	0.894	0.887	0.662	0.834
	DecisionTree	1.218	1.017	0.766	0.848	1.142	0.769	0.889	1.068	1.58	0.874	0.977	0.994	0.836	0.862
	RandomForest	1.02	**0.874**	0.648	**0.809**	1.117	**0.653**	0.614	1.018	**1.499**	**0.727**	**0.847**	**0.85**	0.725	**0.767**
MAE [kW]	CatBoost	**70.16**	**92.12**	100.52	137.99	**125.89**	97.19	49.58	222.49	170.97	105.47	62.46	139.91	98.96	169
	XGBoost	72.43	92.31	100.27	142.36	127.66	99.71	**47.3**	222.82	169.26	**100.97**	69.92	142.69	97.41	167.94
	LightBoost	71.82	94.2	92.45	143.77	126.57	97.18	50.91	227.18	165.77	101.12	66.67	135.23	97.16	166.39
	LSTM	72.04	95.42	**89.85**	**136.27**	127.6	99.12	47.7	**203.85**	208.41	118.33	72.08	141.13	**96.57**	171.84
	DecisionTree	87.46	100.87	107.37	153.22	145.27	113.79	69.7	243.29	168.79	122.23	71.05	155.04	119.76	170.6
	RandomForest	72.46	92.44	94.79	139.87	130.42	**96.68**	50.22	227.59	**164.22**	104.58	**62.24**	**130.59**	108.42	**151.82**
RAE	CatBoost	**1.048**	**0.951**	0.918	1.188	**0.826**	0.883	0.546	0.859	1.488	0.438	0.456	0.832	0.47	0.883
	XGBoost	1.082	0.953	0.916	1.225	0.837	0.906	**0.521**	0.861	1.474	**0.419**	0.51	0.848	0.462	0.878
	LightBoost	1.073	0.972	0.844	1.237	0.83	0.883	0.561	0.877	1.443	0.42	0.487	0.804	0.461	0.87
	LSTM	1.065	0.987	**0.816**	**1.158**	0.833	0.894	0.522	**0.805**	1.793	0.476	0.521	0.84	**0.451**	0.894
	DecisionTree	1.307	1.041	0.981	1.319	0.953	1.034	0.767	0.94	1.469	0.507	0.519	0.922	0.568	0.892
	RandomForest	1.083	0.954	0.866	1.204	0.855	**0.878**	0.553	0.879	**1.43**	0.434	**0.454**	**0.777**	0.514	**0.794**

References

1. Bokde N., Feijoo A., Villanueva D., Kulat K.: A Novel and Alternative Approach for Direct and Indirect Wind-Power Prediction Methods. Energies 2018, 11, 2923; doi:10.3390/en11112923.
2. Kisvari A., Lin Z., Liu X.: Wind power forecasting – A data-driven method along with gated recurrent neural network. Renewable Energy 163 (2021) 1895-1909. https://doi.org/10.1016/j.renene.2020.10.119.
3. Bokde N., Feijoo A., Al-Ansari N., Tao S., Yaseen Z.M.: The Hybridization of Ensemble Empirical Mode Decomposition with Forecasting Models: Application of Short-term Wind Speed and Power Modeling. Energies 2020, 13, 1666; doi:10.3390/en13071666.
4. Zhang J., Yan J., Infield D., Liu Y., Lien F.: Short-term forecasting and uncertainty analysis of wind turbine power based on long short-term memory network and Gaussian mixture model. Applied Energy 241 (2019) 229-244. https://doi.org/10.1016/j.apenergy.2019.03.044.
5. Cadenas E., Rivera W.: Wind speed forecasting in three different regions of Mexico, using a hybrid ARIMA-ANN model. Renewable Energy 2010, 35, 2732-2738.
6. Han S., Qiao Y-H., Yan J., Liu Y-Q., Li l., Wang Z.: Mid-to-long term wind and photovoltaic power generation prediction based on copula function and long short term memory network. Applied Energy 239 (2019) 181-191. https://doi.org/10.1016/j.apenergy.2019.01.193.
7. Hu S., Xiang Y., Zhang H., Xie S., Li J., Gu Ch., Sun W.: Hybrid forecasting method for wind power integrating spatial correlation and corrected numerical weather prediction. Applied Energy 293 (2021) 116951. https://doi.org/10.1016/j.apenergy.2021.116951.
8. Wang Y., Zhou R., Liu F., Zhang L., Liu Q.: A review of wind speed and wind power forecasting with deep neural networks. Applied Energy 304(2021), 117766. https://doi.org/10.1016/j.apenergy.2021.117766.
9. Miao H., Dong D., Huang G., Hu K., Tian Q., Gong Y.: Evaluation of Northern Hemisphere surface wind speed and wind power density in multiple reanalysis datasets. Energy, Volume 200, June 2020, 117382. https://doi.org/10.1016/j.energy.2020.117382.
10. Staffell I., Pfenninger S.: Using bias-corrected reanalysis to simulate current and future wind power output. Energy 114(2016) 1224-1239. http://dx.doi.org/10.1016/j.energy.2016.08.068.
11. Bosilovich, M. G., Lucchesi, R., and Suarez M., 2016: MERRA-2: File Specification. GMAO Office Note No. 9 (Version 1.1), 73 pp, available from http://gmao.gsfc.nasa.gov/pubs/office_notes. (access in October 2021)
12. Lucchesi, R., 2018: File Specification for GEOS FP. GMAO Office Note No. 4 (Version 1.2), 61 pp, available from http://gmao.gsfc.nasa.gov/pubs/office_notes. (access in October 2021)
13. Nahuis, S., Guyeux C., Arcolezi H., Couturier R., Royer R., and Lotufo A. "Long short-term memory for predicting firemen interventions." In: 2019 6th International Conference on Control, Decision and Information Technologies (CoDIT), pp. 1132-1137. IEEE, 2019.
14. Phan Q.T., Wu Y.K., Phan Q.D.: A Hybrid Wind Power Forecasting Model with XGBoost, Data Preprocessing Considering Different NWPs. Applied Sciences, 2021, 11, 1100. https://doi.org/10.3390/app11031100.
15. Zheng H., Wu Y.: A XGBoost Model with Weather Similarity Analysis and Feature Engineering for Short-Term Wind Power Forecasting. Applied Sciences, 2019, 9, 3019, doi:10.3390/app9153019.
16. Friedman J.H.: Greedy function approximation: A gradient boosting machine. Ann. Statist. 29(5), 1189-1232.

17. Chen T., Guestrin C.: XGboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD'16, New York, NY, USA, ACM (2016), 785-794.
18. Patterson, J.; Gibson, A. Deep Learning. A Practitioner's Approach; O'Reilly Media,Inc.: Sebastopol, CA, USA,; pp. 150–158, 2017.
19. Gers, Felix A., Jürgen Schmidhuber, and Fred Cummins. "Learning to forget: Continual prediction with LSTM." Neural computation 12.10 (2000): 2451-2471.
20. Gershman A, Meisels A, Lüke KH, Rokach L, Schclar A, Sturm A. A Decision Tree Based Recommender System. InIICS 2010 Jun 3 (pp. 170-179).
21. Jadhav SD, Channe HP. Efficient recommendation system using decision tree classifier and collaborative filtering. Int. Res. J. Eng. Technol. 2016;3:2113-8.
22. Global Modeling and Assimilation Office (GMAO) (2015), MERRA-2 tavg1_2d_slv_Nx: 2D, 1-hourly, time-averaged, single-level, assimilation, single-level diagnostics V5.12.4 (M2T1NXSLV 5.12.4), greenbelt, MD, USA: Goddard Space Flight Center Distributed Active Archive Center (GSFC DAAC), Accessed 01.02.2021 at doi: 10.5067/VJAFPLI1CSIV.
23. GEOS-5 File specifications variable definition glossary. https://gmao.gsfc.nasa.gov/GMAO_products/documents/GEOS-5_Filespec_Glossary.pdf. (access in November 2021).
24. Botchkarev A.: A new topology design of performance metrics to measure errors in machine learning regression algorithms. Interdisciplinary Journal of Information, Knowledge, and Management, 14, 45-79. https://doi.org/10.28945/4184.