Department of Mechanical and Aerospace Engineering

Statistical models in wind resource assessment and energy yield forecasting

Author: Olatz Uriarte Pérez de Nanclares

Supervisor: David McMillan

A thesis submitted in partial fulfilment for the requirement of the degree

Master of Science

Sustainable Engineering: Renewable Energy Systems and the Environment

2019

Copyright Declaration

Signed: Olatz Uriarte Pérez de Nanclares Date: 21/08/2019

Abstract

Wind power is increasing its production around the world even if its variability and unpredictability make difficult its management and the incorporation of this energy to the grid. For this reason, researchers are focusing on methods to improve measurements which can provide the opportunity to learn about future fluctuations at an early stage in the grid, and handle them in advance. The wind resource assessment can have an important role regarding the energy network, the cost, and its efficiency.

This project focuses on investigating and assessing the reanalysis data available to avoid the cost of the met mast equipment and installation. For this, it is assessed the role that statistical linear models have on wind speed (and other variables), finding the best model and testing it in regions such as United Kingdom, Germany, Scandinavia, France, and Turkey. The model was implemented in R and authenticated in two cross-validation frameworks: k-fold cross-validation and backward cross-validation.

The model formed by the four reanalysis tools (MERRA, ConWX, RVM, and ERA5) and periodicity was the most successful one and it was tested for wind speed in different periods of the year. It was concluded that the regressive model has useful and significant importance on the resource assessment. The results were evaluated by the evaluation metrics RMSE, MAE; BIAS, R2, and correlation. K-fold cross-validation had widely successful performance and backward cross-validation performance depended on the site and the training and testing lengths of the iterations.

The potential of the predictions obtained from the cross-validation frameworks was tested when converted into power. It was concluded that predictions were efficient and could match with site data reducing the costs of resources and avoiding the installation of masts with a further investigation of the study.

Acknowledgements

I would first like to thank David McMillan for being my supervisor and linking me with this project which gave me the opportunity to challenge myself.

Secondly, I would like to thank Fraser Tough for giving the opportunity to carry out this project with RES and for being my advisor.

I would also like to thank Jethro Browell, but especially to Ciaran Gilbert, for the pieces of advice given and the time dedicated to me.

Finally, thanks to my family and friends for supporting me in the hardest moments of this year.

Table of Contents

List of figures

List of tables

Nomenclature

ACF: Algebraic Curve Fitting.

ARIMA: Auto-Regressive Integrated Moving Average.

ARMA: Auto-Regressive Moving Average.

CCF: Cyclical Continuous features.

Cp: Power coefficient.

CV: Cross-validation.

EU: European Union.

LSE: Least-Squares Estimation.

MAE: Mean Absolute Error.

MBA: Mean Bias Error.

MERRA: Modern Era Retrospective-Analysis for Research and Applications.

MET: Meteorological Office.

MRT: Mean of the Reanalysis Tools.

NWP: Numerical Weather Prediction.

R2: Multiple-R-squared.

RES: Renewable Energy Systems.

RMSE: Root Mean Square Error.

SLR: Simple linear regression.

WWEA: World Wind Energy Association.

1. Introduction

1.1. Motivation

Renewable energies have achieved significant importance in modern society as well as a necessary increase in the future. These energies represent the key to attend to the queries of climate change. They can offer alternative economic opportunities and provide access to the electricity supply to people who do not have modern amenities (REN21, 2015). Despite being limited for the moment, the depletion of fossil fuels is allowing this deployment, and they have achieved to be cheaper than other options around the world (Mora, 2008).

Wind energy is an intermittent renewable source which extracts the power of the wind and it turns into electricity. The deployment of wind energy assumed the creation of turbines of different types and shapes in different heights, but the maximum theoretical performance is still to be achieved (Schubel & Crossley, 2014). The reduction in the cost of wind energy production, which achieved to be the cheapest, is attracting more and more researches to focus on this energy.

For this reason, wind energy is, within all the renewable energies, the most developed one. *Table 1* shows the increase of the total wind energy installed capacity per year, the total number and the installed capacity for each country.

Table 1 - Total wind energy installed capacity per year (MW). Source: WWEA

Country/Region	2018	2017	2016	2015
China	221,630	195,730	168,730	148,000
United States	96,363	88,775	82,033	73,867
Germany	59,313	56,190	50,019	45,192
India**	35,017	32,879	28,279	24,759
Spain*	23,031	23,026	23,020	22,987
United Kingdom	20,743	17,852	14,512	13,614
France	15,313	13,760	12,065	10,293
Brazil**	14,490	12,763	10,800	8,715
Canada	12,816	12,239	11,898	11,205
Italy*	10,090	9,700	9,257	8,958
Rest of the World*	91,473	83,473	76,325	67,695
Total general	**600,278**	**546,388**	**489,939**	**435,284**
*Preliminary data **By November 2018				

By the end of 2018, the global capacity of wind turbines reached 600 GW according to World Wind Energy Association (WWEA) (WWEA, 2019). Related to that, wind energy met 14% of the electricity of EU's electricity demand considering not only onshore energy but also offshore energy. This assumed to represent 2% more than in 2017. The countries that more contribution received from wind energy were Denmark (41%), Ireland (28%) and Portugal (24%) (Wind Europe, 2019).

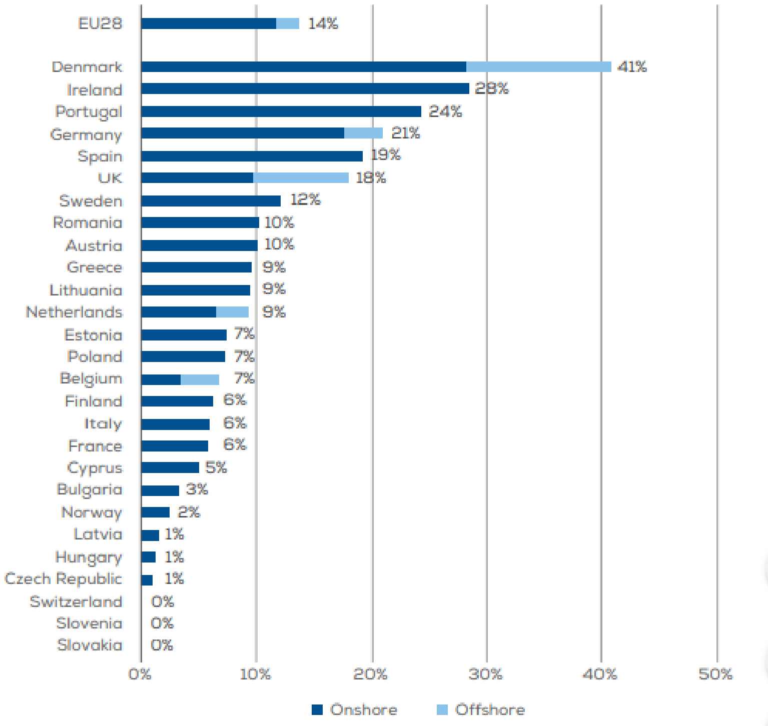


Figure 1 - Percentage of the average annual electricity demand covered by wind (Wind Europe, 2019)

As wind energy is a non-programable generation form and it can be only produced when the wind blows, it can vary even in a very short term, with intermittency possibility and broad change in intervals (Lobo, 2010). Wind energy helps to support in every change of that demand, but the grid needs to ensure the stability in light of variable energy sources (Würth, et al., 2019).

The penetration of wind energy into the grid is growing, so the optimization of diverse methods or parameters/variables are gaining relevance, with the aim to help the turbines capture more wind for electricity generation and gain its reliability (GE Renewable Energy, 2018).

For this reason, it is challenging to forecast in advance and accurately the amount of energy that it is going to produce at each moment. This variability makes especially complex its operation, this is why its future production should be examined, being this affected by an error or uncertainty.

When the wind is reduced, the lack of power must be replaced by other sources with adequate backup power in scale and speed response so that the electricity demand would not be affected.

Other times, it could happen that the total available wind power cannot be embedded into the grid, since wind power is not generated according to the consumption requirements, and it is necessary to reduce the supply of this energy. Therefore, energy resource assessment has become a key issue to make feasible the development and deployment of wind energy and its integration into the power grid (Lobo, 2010).

The volatility of output of wind energy barely allows to the grid take 10% of the wind energy without drastic technical changes or significant costs according to some studies (Beaudin, et al., 2010). Wind power output variability depends on the size of the evaluated site and the wind variations. Small wind farms usually variate hourly more than bigger ones. Variations can be of three types: microscale, mesoscale or macroscale. The first affects regulation (seconds to minutes), the second can affect from minutes to hours and the last from hours to days. The two latter ones can affect an entire region (Beaudin, et al., 2010).

Nowadays, the interconnections with other countries and the linkage with storage sources or hydroelectric also increase the reliability and efficiency (The New Economy, 2018). Hence, some studies are focused on the smoothing effect of the interconnection of wind installations with other regions to reduce the volatility of output, experiencing a direct reaction in the integration costs of the wind energy (Beaudin, et al., 2010).

1.2. Objectives

The project would focus on the reconstruction of past observations of wind speed that would allow assessing the feasibility of the resources in light of replacing the expensive met mast with reanalysis data. The principal statistical method employed is a linear model, varying it in numerous ways to obtain lower errors.

The database is obtained from Renewable Energy Systems (RES) and it is utilised to assess the efficiency of the reanalysis data and develop an appropriate cross-validation framework. It is pursued to obtain the relationship between site data and reanalysis data and generating long term wind speed on-site series. Variable selection will be material for this aim.

1.3. Methodology

First, research is done on the field to know about the state-of-the-art. This idea is outlined in the literature review section. Next, it is decided the tool and language in which the project shall be conducted, and the basic programming commands are learned to interact with the data Renewable Energy System (RES) provided. Next, the familiarisation of statistical models is done, and how these can be implemented in the tool. It is substantial to diagnose the data for the assessment of the quality of this in linear model regressions. Data diagnosis would also help to discover which reanalysis tool is best fitted to the on-site data.

In order to validate the models, a cross-validation framework should be built. Two cross-validations are used: k-fold CV and backward chaining or back-test CV. After this, prediction, training and testing are done, fitting the models on it.

The cross-validation would decide which model is the one that best fits to all the sites available. When analysing the results, comparison of both cross-validations is performed and finally, the predictions are used to see how they fit the on-site data, for the reconstruction of the power output forecast and to see how much potential these would have, using theoretical power curves of the turbines.

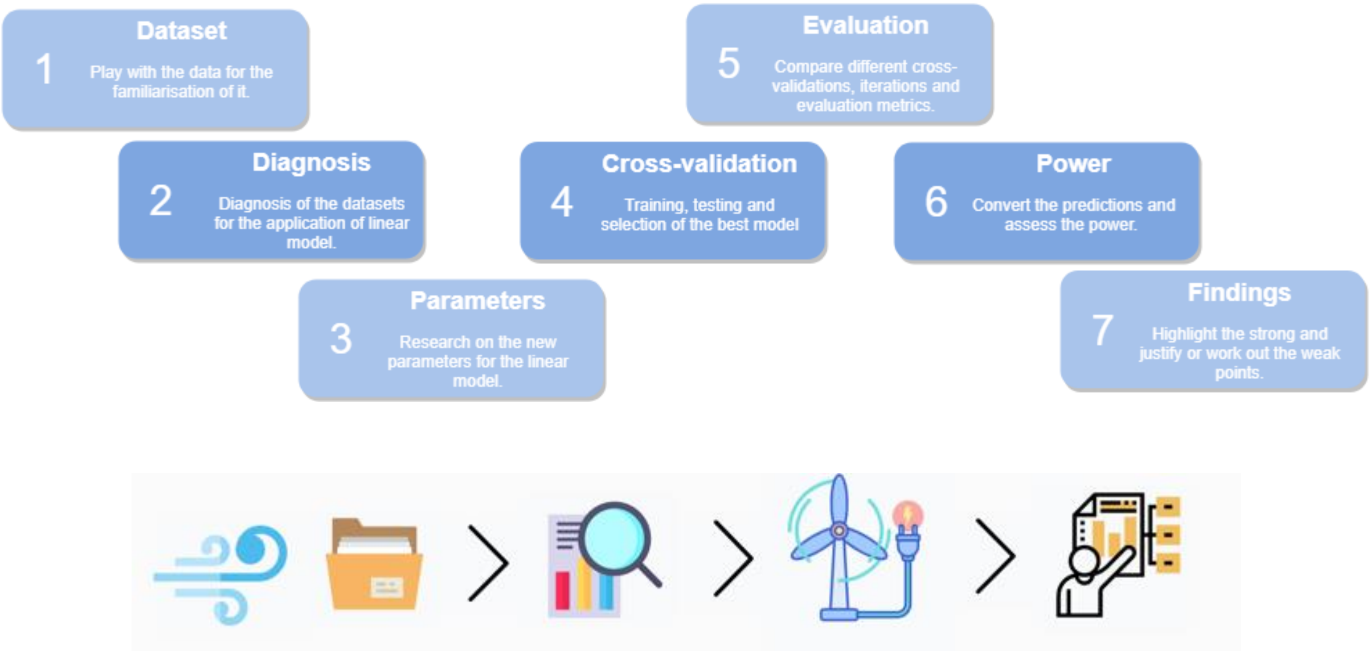


Figure 2 - Representation of the methodology of the project

1.4. Scope

The general goal of this project is to investigate the implementation of the wind resource with the use of a model based on a linear model that fits best any site.

In order to achieve the target, different variables were used to design a linear model that more requirements met. The contribution of all these variables is assessed for the best match in all the sites.

The benchmark for assessing the validity of the model could rely on cross-validation framework. Moreover, two CV's were undertaken to assert the model. These CVs would help to know the error and the effectivity of the model and once the model is selected, this was converted to energy to simulate real conditions.

2. Literature review

2.1. Physical basics of wind energy

The wind is a mass of air in movement and it contains kinetic energy that depends on the speed of the wind, and this is determined by the temperature and the atmospheric pressure. In other words, by the density of the air:

$$E = \frac{1}{2} m_{air} v^2 \tag{1.0}$$

Where, m_{air} is the mass of the air and v is the speed.

The wind direction determines the design of the turbine. Upwind turbines face into the wind while downwind turbines face away (Energy.gov, 2019).

A wind turbine has two/three blades around a rotor which is connected to the main shaft for the spinning of a generator. The turning shaft the kinetic energy of the wind is converted into electricity with the aid of the generator. This electricity is passed through a transformer which allows the use of this being transported into the grid or to a local site (goodenergy, n.d.).

Considering the kinetic energy equation, it can be concluded that the relation between wind speed and the electric power is the energy that goes through the turbine in time.

$$P = \frac{1}{2} \rho_{air} A_r v^3 \tag{1.1}$$

Where P is the electric power generated by the turbine, ρ_{air} is the density of the air, A_r is the area it covers, defined by the diameter of the rotor or the size of the blades and v is the speed.

The equation for the maximum available power allows knowing how power increases with the cube of the wind speed. The power is linear with density and the area.

The available wind speed is measured over a period of time in a specific site before a project is carried out. As wind speed cannot be reduced to zero, a power coefficient Cp is defined with a maximum value of 0.593, which takes the name of Betz limit. The power curve of *Figure 3* shows three different velocities: the cut-in velocity is given when the turbine starts generating power (3 m/s).

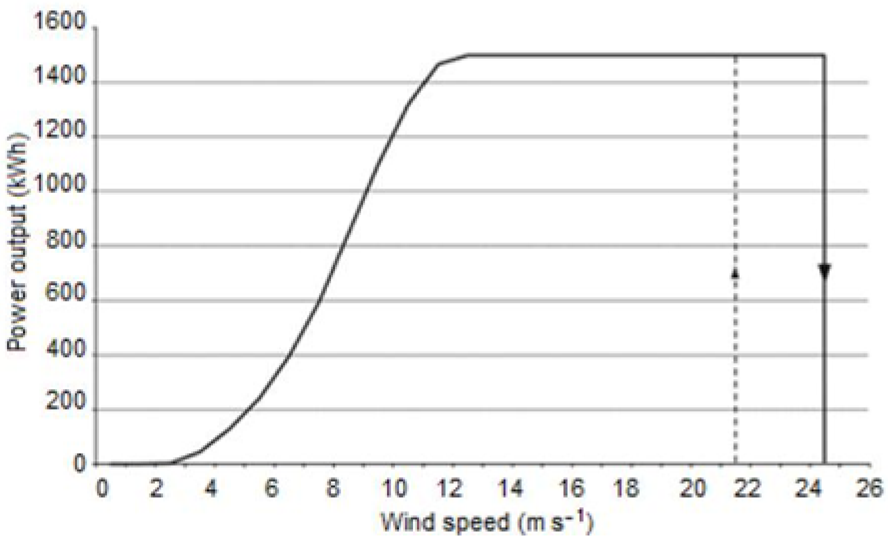


Figure 3 - Power curve of a 1.5 MW wind turbine (Universidad de Chile, n.d.)

The rated speed depends on the rated power and therefore, on the turbine. The rated power is the maximum power allowed by the generator and the control system is responsible for ensuring that this power is not exceeded in high winds (Hansen, 2013). Finally, when the wind achieves the cut-out velocity (25 m/s), the turbines are shut down due to the danger of break (Manwell, et al., 2009).

Within the main characteristics of the wind, the time variability does not allow predicting the output with accuracy. It is a non-stationary process due to the change of the mean and the variance of the wind speed over long time scales. As most of the wind farms are commissioned, it makes more difficult to deal with the variability of the output.

The height of the tower represents a significant characteristic of a wind turbine since the more is the height of the hub, the more is the wind speed obtained, and therefore, the wind power (Hansen, 2013).

2.2. State-of-the-art of wind power assessment

The substantial changes in wind power generation make the integration of this energy to the grid a challenge, as well as for the developer of a wind farm. Wind speed and direction are principally responsible for these drastic changes. Minor changes in these variables can achieve the cut-out speed and lead to a considerable loss of generated power with the shutdown of the wind turbines (Würth, et al., 2019).

Wind power assessment uses wind speed (or more variables) from weather forecasts and on-site real-time measurements for wind power. The developer of a wind farm must do a wind speed measurement and analysis program, providing a prediction of the wind farm’s lifetime energy production (Wind Energy - The Facts, 2019).

For countries that have a high penetration of wind energy into the grid, this prediction is indispensable. It provides the information of how much wind power is expected, when and for how long. The key to this prediction lies in the transformation of the numerical weather data into the power output (Lange & Focken, 2005).

There are five categories in the types of wind energy potential estimation:

- Meteorological potential: available wind resource.
- Site potential: it is based on the previous one but reduced to the geographically available sites.
- Technical potential: calculated from the previous one considering the technology available.
- Economic potential: it is based on the previous one, considering the economic view.
- Implementation potential: it is based on a certain time frame that considers constraints and incentives to asses the capacity of the wind turbine (J. F. Manwell, 2009).

The first efforts for creating forecast algorithms in electric systems were focused on the electricity demand. Demand assessment helped in the decision-making of advance start-up or the condition monitoring of elements of the electric system. But the development of renewables in the last decade made the researchers focus on a new assessment: the production forecast that these energies are going to have. As wind energy is the energy that increased the most in the last years, it is a challenge to transform this energy into a reliable and efficient one.

The value of resource assessment for wind energy in economic terms gains two perspectives. On one hand, the reduction in operational costs produces from the reduction of the necessary reserve. On the other hand, the economic penalisations that the agents have due to the deviations of generation in the electricity market (Lobo, 2010).

2.3. Time series analysis

Forecasting means providing information about the expected variables in a determined time horizon. (Alencar, et al., 2017). They are analysed to interpret the past and predict the future.

This divides time series analysis into four. Very short-term analysis can predict from seconds to minutes and are used for turbine control and load tracking. The short-term analysis is extended to hours and is applied to preload sharing. The medium-term analysis is extended to a day and accepts significant responsibilities with power system management and energy trading. Finally, there is the long-term analysis that can be extended to weeks or years and can monitor a wind farm. Long-term correlation is important for the consideration of variations and seasonal effects. The last two analysis can collect data between 10 minutes and a few hours depending on the predicted period length. Many improvements have been applied to forecast with the use of more input data and the estimation of uncertainties (EMHART, 2018).

2.4. Wind speed assessment

For the wind resource assessment, wind speed/power prediction models have been developed. However, the most recent available measurement is the persistence forecast. It assumes that the wind speed stays the same as in the previous time step (dt) (McMillan, n.d.). It is more accurate than other physical and statistical methods for short-term analysis. Thus, any other method should be compared with this to check the effectiveness of the comparable method (Soman, et al., 2010).

As the value of the wind speed comes from meteorological conditions, historical values of wind speed can be used to predict future values. Numeric Weather Prediction (NWP) is the method used for wind forecasting. Not only the wind speed and direction variables are available in this method, but also pressure, humidity, and temperature (S. M. Lawan, 2014).

Computational programming allows obtaining wind power from weather forecast and on-site real-time measurements. The general assumption in this type of predictions is the stationarity of data. This means that the mean, variance, and autocorrelation structure do not change over time. Non-stationary methods can be turned into stationaries for the use of statistical methods of forecasting. These are typically based on time-series models, such as autoregressive model or neural networks. They allow obtaining a relationship for the differentiation between predicted and actual wind speeds. This forecasting is based on the relation between variables. It is useful since it is easy, cost-effective and provides information timely. Statistical models used, so far, are Algebraic Curve Fitting (ACF), Auto-Regressive Moving Average (ARMA), and ARMA with exogenous inputs (ARMAX), Auto-Regressive Integrated Moving Average (ARIMA), seasonal and fraction ARIMA, among others. Furthermore, neural networks are useful non-linear approaches for wind prediction issues, as it constructs an input/output mapping with interconnected processing units.

Finally, hybrid models also have been developed, which consists of the combination of the two above-mentioned approaches. These models can be evaluated with metrics, such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Bias Error (MBA) (S. M. Lawan, 2014).

2.5. The value of wind resource assessment

As mentioned before, the electricity of the grid needs to cope with the demand at every moment and this is obtained with the balancing of generation and consumption of the electricity.

There are several sources of uncertainty in wind resource assessment, so the accuracy would suppose the improvement of the reliability on the source. The value of a measurement plan is resumed in the aid of the scheduling and anticipation to the future behaviours that would affect the system, managing the necessary resources far enough in advance (Lobo, 2010).

The operation and monitoring of a wind farm require accurate wind measurement technology. The equipment for the measurement and the correct location of a MET mast will ensure the quality of data, but these are cost-effective procedures (Ammonit, 2019). An error of even 3% in the wind speed measurement can also wield a drastic influence on the cost. As weather conditions can vary depending on the season and the year, the interannual variability of the wind can be estimated by a well-correlated long-term reference station, so the uninterrupted wind measurement data gathering can last more than 12 months. The data gaps should be kept to a minimum (less than a week) and it is then processed and evaluated to compare it with long-term meteorological data (NREL, 1997). Longer measurements reduce the uncertainties and they can have a more accurate perception of the wind in the area (Miceli, 2017).

Within the measurement parameters, there are:

- Equipment type, quality, and cost
- Number and location of monitoring stations
- Sensor measurement heights
- Minimum measurement accuracy, duration, and data recovery
- Data sampling and recording intervals
- Data storage format
- Data handling and processing procedures
- Quality control measures
- Format of data reports (NREL, 1997)

Apart from the wind speed, wind direction is also crucial for this measurement; it avoids sheltering effects in wind farms. Other variables can significantly improve the performance of the measurement, such as air density, air pressure and humidity (Ammonit, 2019).

The use of a wind index can help to estimate the energy production with the historical record of mean speed. Trend information would help to predict the lifetime of a wind farm, which can last between 20 and 30 years. Climate can change from what was assumed at the beginning of the plan.

Meteorological variables predictions are converted to power predictions using the power curve. Statistical methods help to estimate the conversion function, which is non-linear, bounded and non-stationary due to the time period or environmental situation (Pierre Pinson, 2008).

The components for a measuring system depend on climatic and regional conditions, as well as, the size of the wind farm. The height of the measuring tower is dependent on the turbine hub height. The hub height will set the wind performance and this the return on investment (Ammonit, 2019).

The horizontal values of wind speed have a different value depending on the height, which adopts the name of wind shear. The higher the height the lower the friction this would have. So, adapting the model to the time series of wind speed, the wind speed must be scaled to an electrical output. Typically, the wind shear component (hub height scaling factor) is 1/7. This depends on the roughness and atmospheric stability, and this value is specifically applicable to a low surface and well-exposed site. It is converted following the next equation (Dean Laslett, 2016):

$$\frac{v}{v_{ref}} = \left(\frac{z}{z_{ref}}\right)^{\propto} \qquad (1.4)$$

Where, v is the wind speed at height h, v_{ref}: is the wind speed at MET station, z is the typical turbine hub height and z_{ref}: is the typical MET station height.

2.6. Uncertainties

As it is not possible the exact reconstruction of observations of wind energy of a wind farm, it is practical to determine or give information about the uncertainties of prediction. This would help to know the distribution of error and handle the production of the energy in an optimum way.

Among the wind uncertainties, there is the instrument accuracy, the measurement period, long-term wind data correlations, historical wind potential, future wind potential, and wind flow modelling. Also, the consistency of reference sources that determine the extent to which the wind measurements are consistent and the systemic errors or other external factors with which the measurements are influenced by. Finally, there is the wake modelling due to the complexity of inter-turbine impact and the climatic variability (the trend that wind speed will have in the future) (Hawker, 2019).

Uncertainties can be reduced providing to the evaluation metrics an estimation of the error associated with the prediction, these are the prediction intervals, which limit the estimation of the probability of real production within the generation band expected. These prediction intervals are frequently used in statistical literature.

The location of the mast is a crucial variable of a wind farm performance. Turbulence, terrain inclination, and vertical wind component are the primary parameters when examining the decision of location, conditioning the wind regime of the site. Every installation requires previous wind analysis (Kalkan, 2015).

Within the energy uncertainties, selecting a power curve to obtain final power prediction is a controversial issue since each turbine has a different power curve, that unfortunately often go hand in hand with. The location of the masts, the location of the reanalysis nodes can also affect the accuracy of resource assessment (I.González-Aparicio & A.Zucker, 2015).

3. Reanalysis products

Reanalysis is a reproduction of consistent and systematic meteorological data using a few observations in a specific time of period, for example, hourly. Reanalysis products create global datasets that are based on the historical data of the atmosphere, land surface and oceans. They are used for monitoring climate change. The reanalysis has developed during the centuries achieving a significantly improved spatiotemporal resolution (Hyun-Goo Kim, 2018).

The data from RES is measured with the tools RVM, MERRA2, ConWX, and ERA5.

3.1. MERRA

MERRA (Modern Era Retrospective-Analysis for Research and Applications) is a tool that allows obtaining global reanalysis datasets from meteorological models. Available data is from 1979 on and provides information of variables, units and data files (NASA, n.d.).

Examples of the parameters that can be downloaded are wind speed at different heights/pressure levels, wind direction, displacement height, temperature, moisture content, air pressure (Olauson & Bergkvist, 2015).

The basic approach of this only uses wind speed and direction but it has many other variables to use, such as pressure, temperature, relative humidity, time of the year and time of the day (Olauson & Bergkvist, 2015).

3.2. ERA5

ERA5 provides hourly information about atmospheric, land and oceanic climate variables. The data cover the Earth on a 30-km grid, and it provides uncertainties for variables at reduced spatial and temporal resolutions.

3.3. ConWX

ConWX provides weather analysis up for any site or region, with different variables such as the wind, temperature, and precipitation. It provides the essential information for the wind power forecasting with any height wind speeds ensuring maximum accuracy. It can also provide parameters that influence energy pricing and trading decisions (ConWX, 2019).

3.4. RVM

No information found.

Table 2 - Reanalysis Tool's features

Reanalysis dataset	Institution	Resolution (°lat x °lon)	Vertical level (m)	Time resolution (h)
MERRA2	NASA	1/2 x 2/3	50	1 (time averaged)
RVM	-	-	80	1 (time-averaged)
ConWX	ConWX	(1-10 km)	50	1 (time-averaged)
ERA5	ECMWF	0.25 x 0.25	100	1 (time averaged)

Each reanalysis tool has nodes of measurement around Europe. These nodes could be closer or far away from the masts to assess. It is assumed that the less the distance the better the accuracy predictions will achieve, and a relationship between these and the site data should be captured for the evaluation. In the following figure (*Figure 4*), there is a visual representation of a proposed site mast measurement in yellow and the reanalysis tool measurement point location in red or orange.

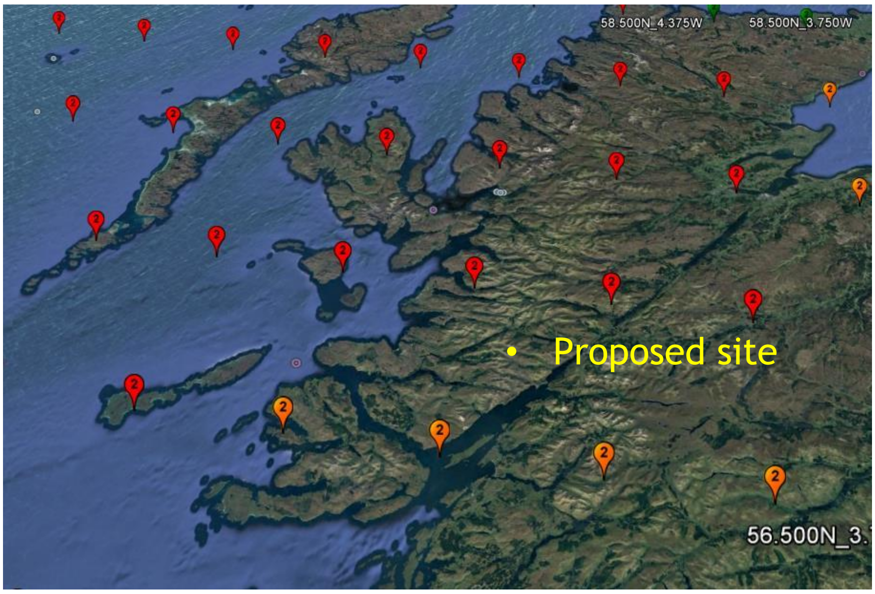


Figure 4 - Visual representation of nodes and site measurement location (Technical projects in the renewables industry, 2019).

4. Methods

Once the basic principles of the literature review are assimilated, data diagnosis would allow knowing the patterns of the data for the use of statistical models afterwards. Data diagnosis would also help to discover which reanalysis tool is best fitted to the on-site data. Then the statistical models are explained. In order to validate the models, a cross-validation framework should be built. After this, prediction, training and testing are done, fitting the models on it.

Every programming part of the project is implemented in language R and using RStudio.

4.1. Statistical approach

Statistical prediction consists of the analysis of the values of a variable to compare it with other variable values to find significant patterns that would conclude in the knowledge of the value that this variable would take over a period of time.

The regression analysis has a first step which consists of the data collection process and it will be then based on that data. The data collection methods can be a retrospective study based on historical data, and observational study or a designed experiment. An accurate data collection can end in a more applicable model. There are a lot of forms of regression, but each form has its condition where they are best suited to. Regression analysis is a form of predictive modelling technique and can be used for data description, parameter estimation, prediction and estimation and control (Peck & Vining, 2012).

It obtains a relationship between a dependent and independent variable, as well as the strength of the impact of independent variables on a dependent one. It is used for forecasting and time series modelling. It is a tool for modelling and analysing data and compare the effects that a variable has on different scales.

There are two types of implementation considering the regression. First, the linear model and sophisticated splines, where quadratic and cubic splines are analysed.

Data Regression Diagnosis

To begin with regression models, four assumptions associated with the linear regression model should be assessed.

- Linearity: X and the mean of Y have a linear relationship.
- Homoscedasticity: X values have the same variance of residual.
- Independence: Observations are independent of each other. Ensuring the independence of the responses of y is one of the aims of data collection. As the responses have the same source, they can present dependency (Peter K. Dunn, 2013).
- Normality: Y is normally distributed for X values. F- and t-tests are used for the evaluation of this assumption. Even if the residuals are not normally distributed, these two tests have good responses if the number of observations is large and the outliers are not serious. This assumption is more critical when applying to small size data (Peter K. Dunn, 2013).

First, a histogram is used to see the distribution of the data. This would show whether the data is normally distributed.

Applying a simple linear regression, the output in R provides a brief numerical summary of the residuals, estimated regression results, and statistical tests. The output provides the value of r-squared which explains the variation of the response variable Y is explained by predictors, and not by error (James, et al., 2013)

The basic tool for the examination of the fitness of a model are the residuals. In R, using the command *which=1:4*, the first plot describes the residuals versus the fitted values, being the residual values the subtraction between the observed y and y model predicted. This plot can help to assess linearity and homoscedasticity. The assumption is not useful if the residuals obtained are very large with big positive and negative values (people.bu.edu, 2019). The second plot shows the standardized residuals versus theoretical quantiles. Keeping with the assumption of linearity, the residuals could not be far away from 0. This is called the QQ-plot and can also affirm the normality assumption whether the observations chart a 45-degree line. It is expected to obtain a straight line if the data come from a normal distribution with any mean and standard deviation. The third plot assesses the homoscedasticity. Square rooted standardized residual vs. predicted value plot is represented, and this cannot prove a pattern since they would be distributed evenly across the line y = 0 (James, et al., 2013). Finally, the fourth plot is Cook's distance plot, which provides the information of how much does the estimation model change with each observation. Any observation should not be close to 1 or more (people.bu.edu, 2019).

Simple linear regression (SLR)

An autoregressive model predicts future behaviour based on past behaviour, in other words, it is the product of past observation. Its coefficients are calculated by least squares regression. It is a type of linear regression based on historical data. The formula for simple linear regression is:

$$Y_t = \beta_0 + \beta_1 x + \varepsilon \qquad t = 1,2,\dots,n \tag{4.0}$$

Y_t is the response variable at some point in t directly related to X_t the predictor variable. Autoregression model differs from linear regression because apart from Y being dependant on X, it is also dependent on previous values of Y. β_0, the intercept, β_1 the slope, and ε_t, a random error component, are unknown values that are taking values depending on the data owned.

After fitting a linear regression model, it should be determined how well the model fits the data. It must be carefully checked if there is a strong relationship between the data (correlation) and if the relationship is linear enough to be approximated by a straight line. The data used to fit the model parameters (β_0 and β_1) is representative of the range of normal operating conditions and that the assumptions regarding the residual hold (Peck & Vining, 2012).

Linear regression helps to identify which is the smallest difference between the observed and fitted values, that is, it identifies the sum of squared residuals (Frost, 2018).

Least-Squares Estimation

β_0 and β_1 are the regression coefficients and they can be interpreted with the Least-Squares Estimation (LSE). This means according to (Peck & Vining, 2012) that 'the sum of the squares of the differences between the observations y_t and the straight line is minimum'.

$$S(\beta_0, \beta_1) = \sum_{t=1}^{n} (y_t - \beta_0 - \beta_1 x_t)^2 \qquad (4.1)$$

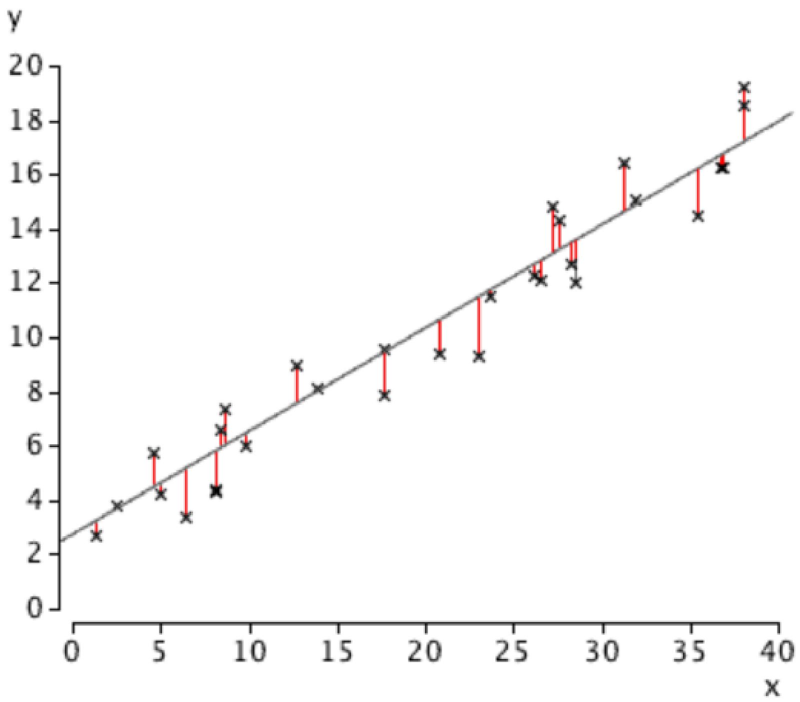


Figure 5 - Minimise sum of squares (Cast Massey University, 2017)

Figure 5 provides an explanation of *Equation 4.1*. The vertical distances from the crosses to the line, the residuals, indicate the matching between predictions and actual data.

Non-linear Regression

When the variables have not a linear relationship between them, it is suggested to shift the regression to a polynomial model. This could have different degrees. When the degree is 2 is called quadratic, 3 cubic, 4 quartic and so on.

Polynomial models allow having other predictor variables by raising each of the original predictors to a power. The standard method to extend linear regression is to replace the linear model with a polynomial function. Polynomial regression is sometimes difficult to handle due to the number of features that appear with the complexity of the formula and it has an over-fit tendency (Singh, 2018).

Cubic spline regression gives more flexibility to the data to be fitted. A few knots should be selected for the usage of this regression. They have been used in the prediction of wind turbine power and power curve modelling (Shahab Shokrzadeh, 2014).

Multiple linear regression (MLR) uses explanatory variables for the prediction of the principal variable. It is an extension of the SLR. The independent variables cannot be highly correlated between each other (Kenton, 2019).

4.2. Evaluation metrics

The evaluation of the models is essential for any project. A model can provide different results if measured with different metrics, so more than one should be used for the assurance of good performance and accuracy (Mishra, 2018). According to (Swalin, 2018), MAE and RMSE are the two most popular metrics in regression.

MAE

Mean Absolute Error is the average of the difference between observations and predicted values. It shows the positive error average. However, this metric does not provide information about the under or overprediction of the data (Mishra, 2018).

$$MAE = \frac{1}{N}\sum_{j=1}^{N}\left|y_j - \hat{y}_j\right| \tag{4.2}$$

- N is the number of samples
- y_j the j predicted the value
- $\hat{y}_j$ the j measured value

RMSE

Root mean squared error is the index most often used to measure prediction accuracy (McMillan, n.d.). It is the average deviation of the residuals. Residuals are the differences between the observed and the predicted values. It is useful in climatology, forecasting, and regression analysis (Glen, 2016).

$$RMSE = \sqrt{\frac{\Sigma(y_j - \hat{y}_j)}{N}} \tag{4.3}$$

Parameters are the same as in MAE. These results must be higher than MAE error values.

This gives information about the observed unknown outcome and the predicted ones from the model.

BIAS

Bias provides the high or low results of estimation compared to actual values. The value of bias is the mean of the difference between the estimated and the actual values. Bias can be a positive or negative error (Smith, 2017).

$$BIAS = \frac{1}{N}(y_j - \hat{y_j}) \qquad (4.4)$$

Parameters are the same as in MAE and RMSE.

R-squared and Correlation

R2 (multiple-R-squared) value indicates how much variation (percentage) is captured by the model. The closer to 1 the R2, the larger value of the variance of the model is explained, and hence, a good fit. As it is squared, it can never be negative.

Correlation, instead, is the degree of relationship between the two variables and it can be between -1 and 1, having the meaning of the unison movement of the variables (Glen, 2016).

5. Model validation - Cross-validation framework

Cross-validation (CV) framework is used to verify the skills of the machine learning models used. It compares and selects a model for predictive problems since it estimates lower trends. It is straightforward to understand, to implement and compare with other methods. It is used when the forecasting is the final objective and it is wanted to estimate the precision of a model that is going to put into practice (Brownlee, 2018).

5.1. K-fold Cross-Validation

k-fold Cross-Validation consists of the data partition into two datasets. The first is the training data (and validation data) and the second is the test set. The training set is used to train the model and it can vary. Validation data set is used to run the model trained in the training set. It provides a cogent evaluation of the model's fitness. Finally, the test dataset is used for the evaluation of the final model, being this between 5% and 20% of the dataset.

Once training data is divided into k folds, one-fold is used as a training data and the other as validation. This is repeated k times changing the testing fold in each iteration. Once the iterations are finalised, the precision and the error of each model is calculated with the mean of the training models.

Obtained the mean accuracy for the mode, cross-validation process can be repeated for the other classification methods and the one which has the best values and lowest error is chosen. Finally, this model can be applied in the validation set mentioned before, since it is assumed that this model is the best result obtained during the training phase (Delgado, 2018).

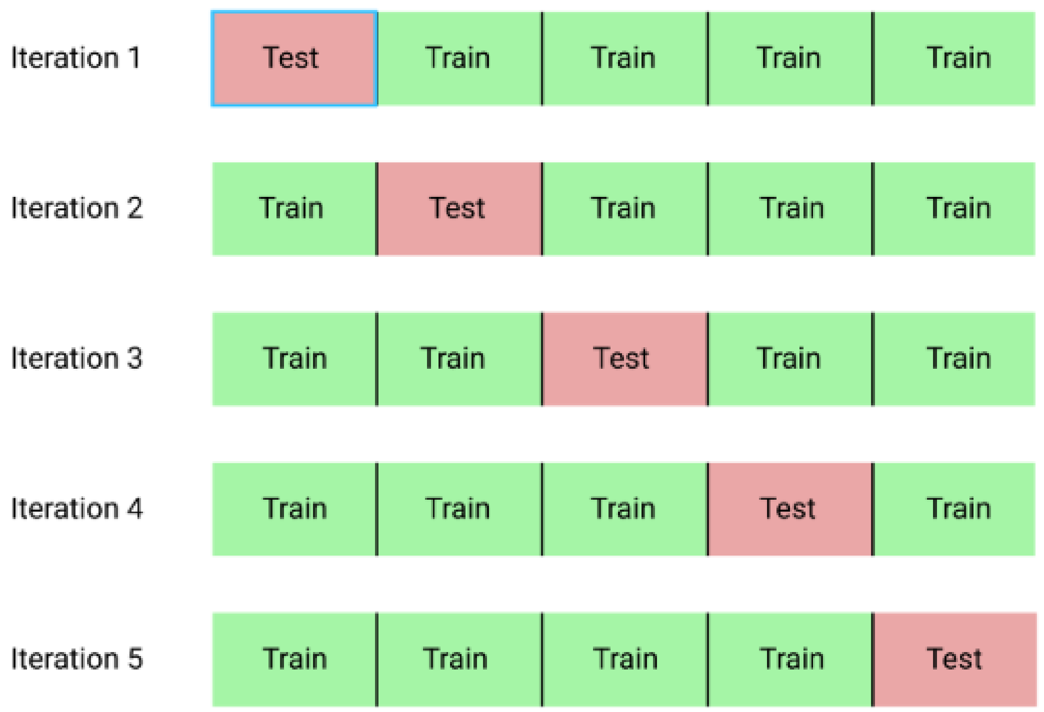


Figure 6 - k-fold cross-validation (Shaikh, 2018)

5.2. Backward/Forward chaining cross-validation

Backward chaining CV (BCV) is used for time series models since it is conditioned by temporal dependencies. In this CV, datasets are split again, given a different length to the training window and before the testing window. The testing window is located before the training window, in other words, the training set only considerate the observations that occurred before that observation. So, it consists of a single observation. As the training and test sets are chosen randomly, each iteration can obtain diverse results (Hyndman, 2016). Many numbers of samples are carried out for the better efficiency of the model.

As represented in *Figure 7*, the training set is the one in which MERRA wind speed and on-site wind speed are represented. The model is obtained for that window and tested in the previous set where only MERRA wind speed is represented. Predictions are obtained for this set, obtaining on-site data for the whole graph, represented in red as Hindcast on-site wind speed.

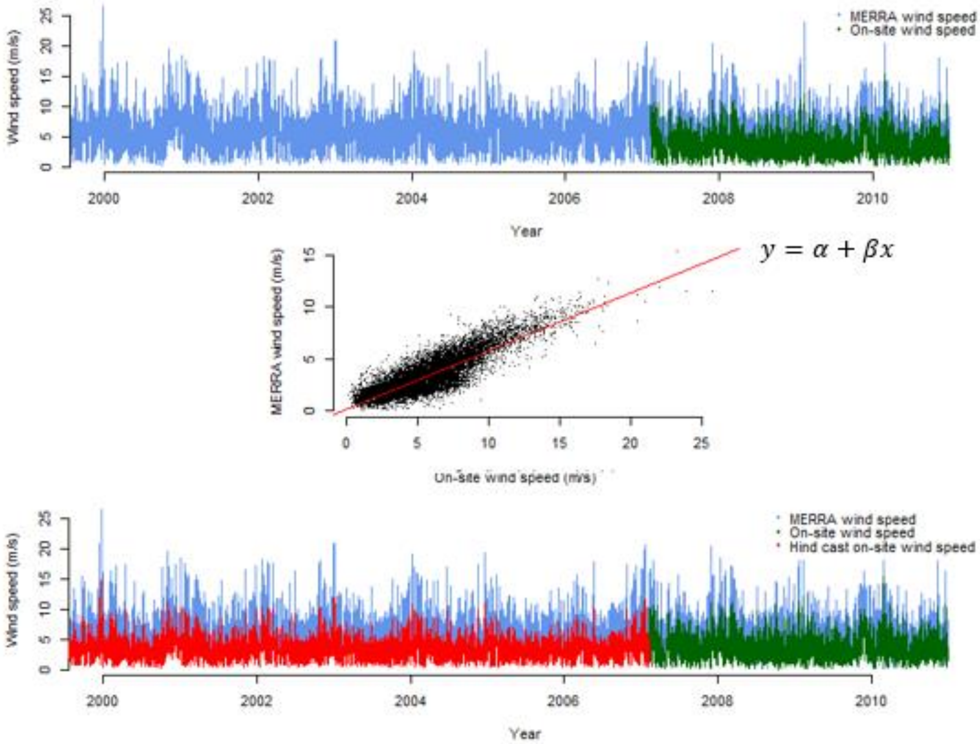


Figure 7 - Backward/forward chaining cross-validation (Technical projects in the renewables industry, 2019)

There are two types of windows in cross-validation, sliding window and expanding-window. It can be stated that in the two CV used both are used. In k-fold CV, sliding window, as the same training size is maintained and the window slides across the data. In Backward CV, expanding-window, the training set is random, and it is expanded or narrowed depending on the iteration.

6. Data

In this project, the observations of the wind are done with different data sources. Surface wind speed from anemometer recording is available at the hourly frequency and the data is extended between 2010 and 2016. In the following table, the location of each mast is represented.

Table 3 - Location of the masts

MAST	REGION	LATITUDE	LONGITUDE
ENGgraM277	England	52.750507	0.166926
ENGdenM224		50.786999	-3.860358
FRAotrM510	France	47.841888	3.89383
FRApdlM29		43.68219	3.89383
GERlacM1	Germany	48.32077	8.189299
GERlirM1		48.029728	8.247731
GERpr1M2		48.22121	8.167225
GERpruM1		50.284874	6.385167
NIRccgM352	Northern Ireland	54.62513	-7.579291
NIRcgrM78		54.99501	-6.805172
NIRnidM2		54.805318	-6.063413
NORskvM30	Norway	58.646568	7.383308
NORvarM32		58.833626	5.905409
SCOdunM103	Scotland	57.225857	-4.274011
SCOfreM73		55.77282	-5.461154
SWErodM39	Sweden	57.986187	12.499485
TURcigM41	Turkey	38.571198	30.256189
TURevrM105		41.642799	27.694426
TURhvzM20		40.955791	35.558552
WALbryM578	Wales	52.011795	-4.173964
WALglhM294		52.414738	-3.276726
WALmmrM281		51.709217	-3.794198

Doing this table more visual, the sites are represented in maps which will help in the error interpretation and comparison later in this document.

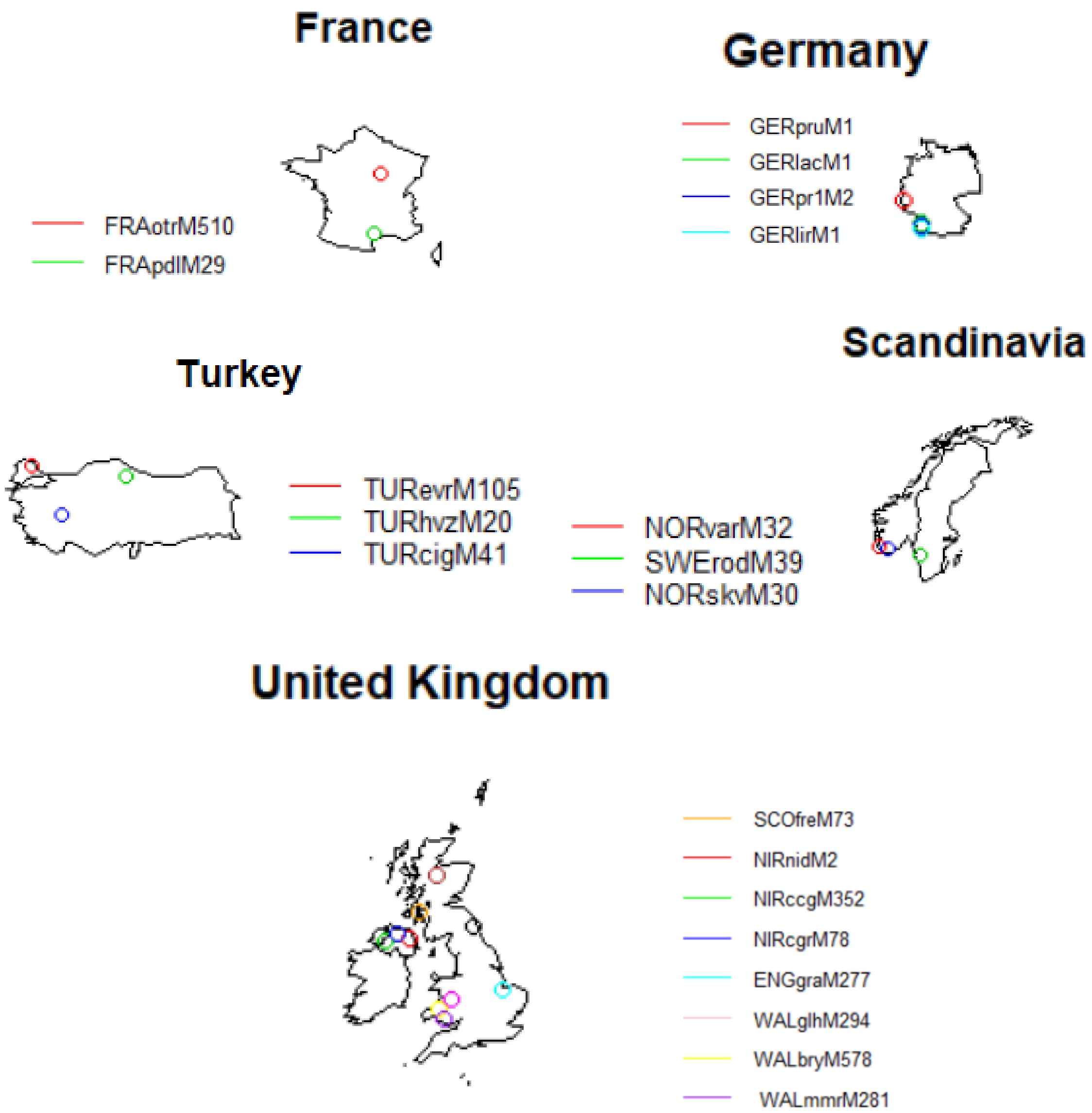


Figure 8 - Visual location of masts

It should be considered that between these regions the United Kingdom is the region that strongest wind (21,4 km/h in January and 15,5 km/h in August) has. This has its softest months between April and October and the wind with increased frequency is given between April and May. This is followed by Germany (19,4 km/h in January and 14,3 km/h in August), which has its softest months between April and November.

Then comes Turkey (19,2 km/h in January and 13,8 km/h in August), which changes the pattern a bit having its softest months between March and July, and the increased frequency wind is given between September and November. France (18,8 km/h in January and 12,6 km/h in August) has its softest months between April and October and the wind with increased frequency is given between April and May. Finally, Scandinavian winds (17,7 km/h in January and 9,1 km/h in August) are the softest ones between April and October and the increased frequency months are between April and June. All this data should be considered when analysing the results (WeatherSpark, 2019).

The first step when analysing the data is the identification of the variables it has. Data is complemented by the wind speeds at the site and the wind speeds of reanalysis tools. With this information, a model of relationships can be built between two different wind speed variables. In this way, a 'normal behaviour model' can be established.

It is also important to analyse the length of the data and how far the reanalysis measurement points are from the site anemometer. This can be easily interpreted in *Table 4*.

Table 4 - Data length and distance to the mast

MAST	Length (hours)	Length (years)	Mast to MERRA distance (km)	Mast to RVM distance (km)	Mast to ConWX distance (km)	Mast to ERA5 distance (km)
ENGgraM277	14481	1.7	24.9	0.8	1.1	10.1
ENGdenM224	8476	1.0	25.8	0.2	1.2	4.1
FRAotrM510	56073	6.4	20.6	0.4	0.4	15.8
FRApdlM29	30320	3.5	21.4	0.3	1.6	14.9
GERlacM1	9780	1.1	20.5	1.4	3.8	7.0
GERlirM1	8440	1.0	9.7	2.3	1.2	11.5
GERpr1M2	9598	1.1	24.8	0.6	1.4	10.1
GERpruM1	6720	0.8	25.8	0.7	1.7	14.2
NIRccgM352	47693	5.4	14.8	0.5	0.8	5.8
NIRcgrM78	32756	3.7	4.5	0.6	1.9	12.2
NIRnidM2	45455	5.2	24.7	1.0	0.7	11.3
NORskvM30	8328	1.0	17.7	4.6	1.1	17.7
NORvarM32	7871	0.9	27.1	5.3	1.5	6.6
SCOdunM103	13570	1.5	25.9	0.5	3.8	9.4
SCOfreM73	35596	4.1	27.3	1.4	1.5	4.9
SWErodM39	22728	2.6	1.5	3.1	0.7	11.3
TURcigM41	26944	3.1	23.7	1.0	0.4	14.8
TURevrM105	28004	3.2	22.7	1.5	1.9	10.1
TURhvzM20	17782	2.0	7.4	0.7	19.8	20.0
WALbryM578	35279	4.0	13.8	1.0	0.5	12.6
WALglhM294	29915	3.4	14.0	1.7	1.7	9.6
WALmmrM281	10653	1.2	23.5	0.4	0.3	14.2

Before applying any model, all the reanalysis data is compared with the site data in a time series to see if there can be a relationship between them. After doing the data diagnosis, the best option to do this is through a selection of the data limiting it to the first 100 points. This would do the visualisation easier to analyse. Although the figures and graphs that are shown after in this document are not for every site, each step explained is done with every site and reanalysis data. It should be highlighted that R uses ordinary squares method to fit the linear model.

6.1. Application of the linear regression in data

First, different models are explained. As wind speed is the only variable available in the datasets, new variables should be created for the implementation of the model. The selection of parameters is crucial for the best performance of the resource assessment.

The models used are the following ones:

- Simple linear regression (SLR)
- Polynomial regression
- Cubic spline regression
- Multiple linear regression
    - Encoding cyclical continuous features (CCF) - 24-hour time

$$\sin \frac{hour * 2\pi}{24} \tag{5.0}$$

$$\cos \frac{hour * 2\pi}{24} \tag{5.1}$$

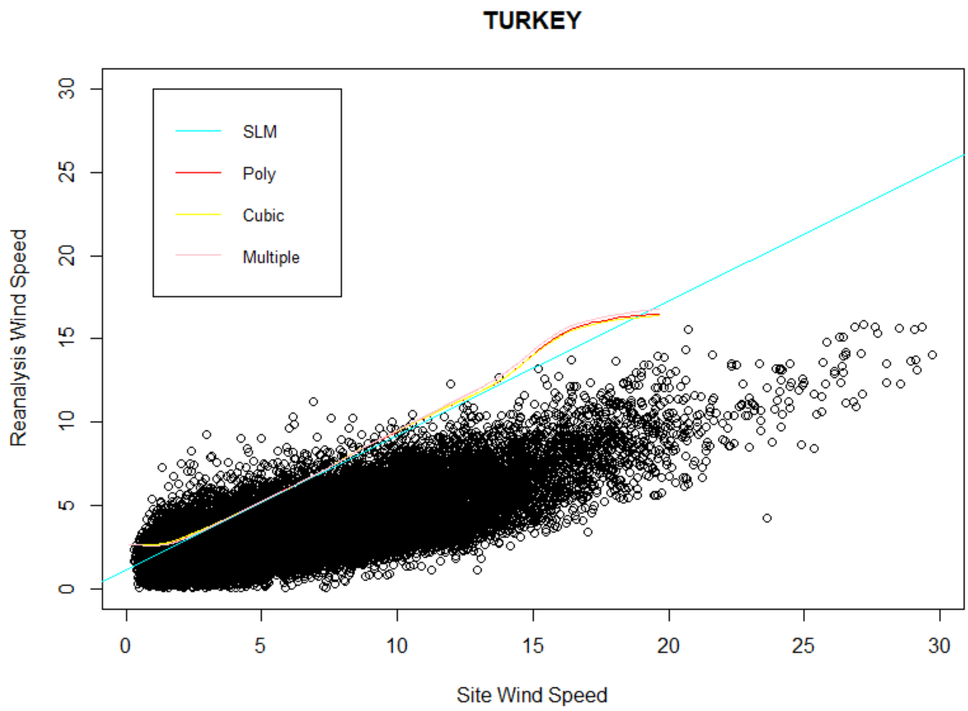
Previous formulas are also assessed with 4 π and 6 π.

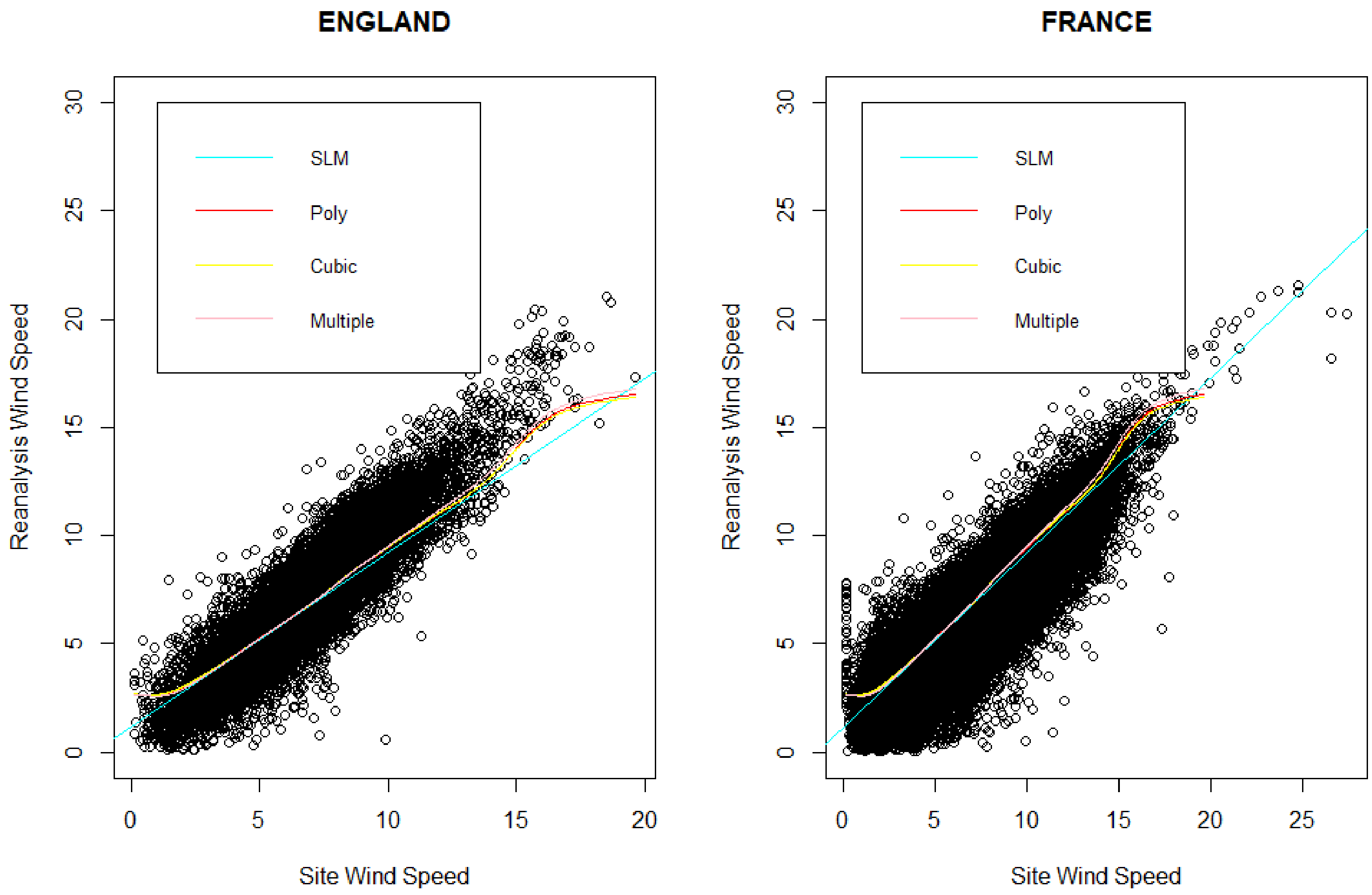
Also, different slopes and intercepts for hours and months:

- Multiple linear regression with wind speed multiplied by the hour
- Multiple linear regression with wind speed multiplied by the month

Variables modified for these models were every reanalysis data, using some of them and, obtaining the difference between them. In the next table, this can be easily understood.

Formulas from [Appendix I](#) are formulas already shifted. As some of the formulas had poor results from the beginning, they were directly discarded, such as the slopes an intercept of month and year. It should be highlighted that not every model is represented in the document.




ENGLAND
FRANCE
SLM
Poly
Cubic
Multiple
Reanalysis Wind Speed
Site Wind Speed
0
5
10
15
20
25
30

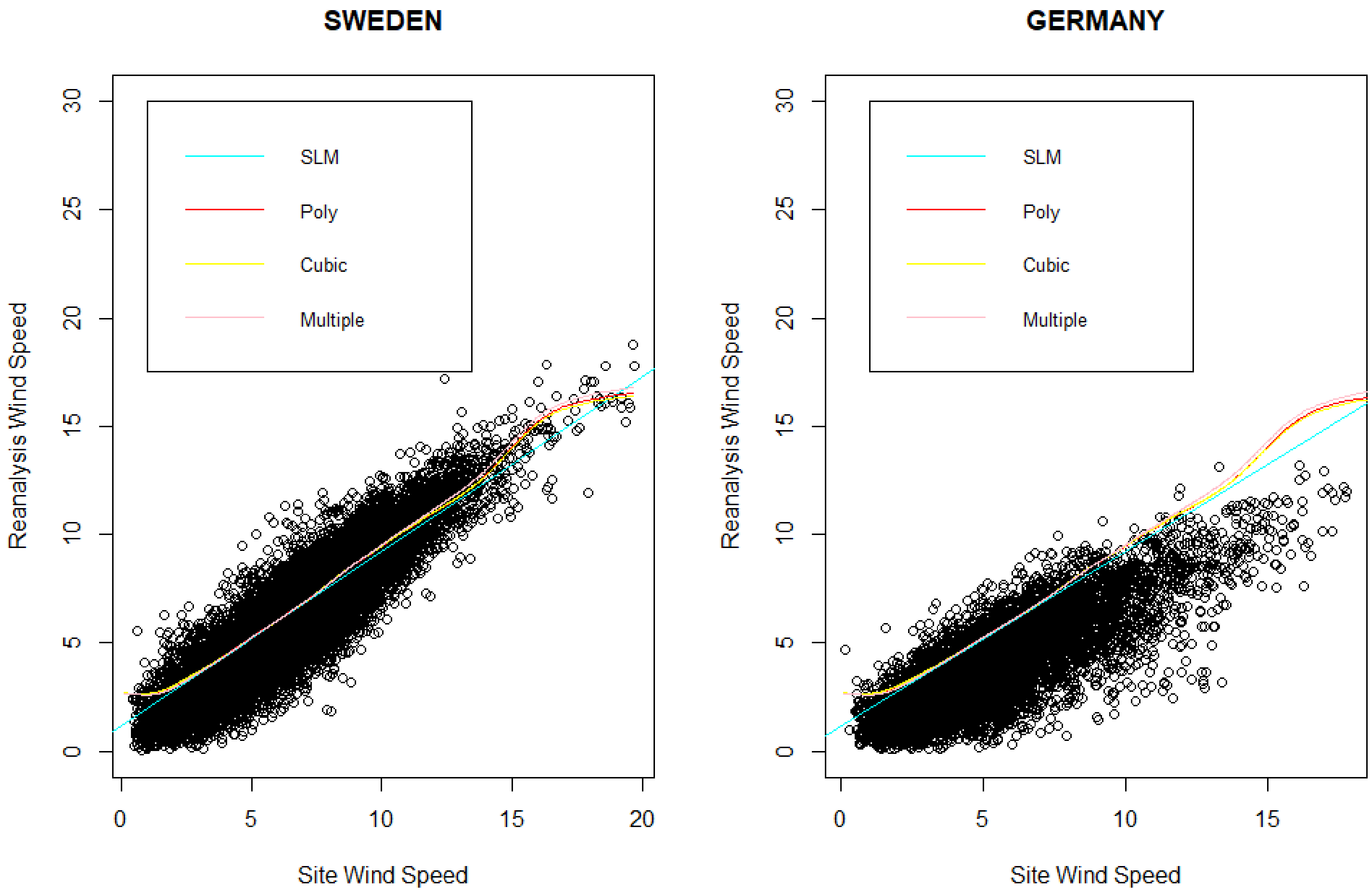


Figure 9 - Comparison between best models and a simple linear model for each region

In *Figure 9*, it is done a comparison between the simple linear model, polynomial model, cubic model, and multiple linear models. In this way, the flexibility that a linear model can give is proven, only with the use of new variables. This is done for each region available and it can be detected that Germany's and Turkey's data is more scattered, so those two regions are more difficult to adjust to the linear model than the others. This should be considered in later results.

7. Analysis and Results

In this section, results for different cross-validations, training lengths and folds were calculated for all the available datasets.

7.1. k-fold Cross-Validation testing

k-fold CV was created splitting every site data into folds. Data was trained in a fold and tested in the rest, creating in this way a k iteration process. Once that all the iterations were done, the mean of all the results was worked out. Here, the training length of the data depends on the number of folds chosen.

All the models from *Appendix I* are tested in the k-fold cross-validation framework, which obtains a BIAS, RMSE, MAE, correlation and the R2 values for each fold, each model and each site. As mentioned, the mean is obtained for every evaluation metric of each dataset. The rise of folds was done with 5, 6 and 10 folds and the later one is assumed to obtain better results in the datasheets with larger lengths.

Model selection

At this point, a decision should be reached for the shifting of models. All the reanalysis tools were assessed individually but the interactions between them provided better results. The coefficients that variables were assigned in the linear model can provide a useful tip about the relative importance of the variable. Generally, the higher the value (absolute value) of the standardised regression coefficient, the higher the weight (importance) a variable is assigned. For this reason, every model was assessed, and the coefficients of the reanalysis tools in the regression model indicated that ERA5 is the tool that more importance had. This is followed by a conflict between ConWX and MERRA, and the last position was given to RVM.

The behaviour of the reanalysis tools, globally, was based on the height of the mast and the distance from the measurement points to the mast. ERA5 was one of the best models in every site since it obtained great contribution irrespective of the height or distance. It was considered that MERRA and ERA5 were the reanalysis tools that further away were from the masts in the analysis. However, other reanalysis tool equaled or overcame ERA5's contribution when they had characteristics that supported them. ConWX was the only reanalysis tool that could overcome ERA5 in some models, which best contribution was given when the mast's height was between 50 and 70.

There was no pattern found for MERRA, but RVM had a good contribution in those sites in which the mast had a similar height to 80 m (the measurement height of RVM). It only failed this pattern in NORvar, where the distance to the mast distanced to 5.3 km.

Once this was carried out, since the gap between the error of the different models tested was not wide, a For loop was created for the choice of best models. Models were shifted by the minimum error in every evaluation metric and the maximum R2 and correlation. In this way, it was concluded that BIAS was an ambiguous metric that did not help in the decision-making of the models but RMSE and MAE did. The models that achieved the lowest BIAS did not coincide with the ones that had lower RMSE and MAE. However, all the sites coincided that the most successful models were the ones that considered the four-reanalysis tool in the datasheets. Also, the periodicities were added to increase the chance to succeed.

Principally, 5-, 6- and 10-fold did not coincide in the lowest BIAS model, since this was given almost once for every model. Even though, model 29 was the model that most times achieve this lowest value. When it comes to speaking about RMSE and MAE, models 35, 36, 41 and 45 performed better. Correlation agreed that models 35 and 36 were the best for every fold, although in other models were also given high values. Nevertheless, there was no doubt when considering R2, since all the sites coincided that model 36 had consistently the highest value.

5-fold cross-validation

In *Appendix II*, all the results are represented for each analysis. First, the models 35 and 36 were compared in the 5-fold CV as shown in 5-fold cross-validation. Models did not show a large difference between their evaluation metrics, but, as mentioned before, model 36 was the model that best fits, due to it was the model that most of the times had the highest R2 and correlation most of the times. Below this table, the worst case of the linear models is represented, this was given by model 33 (variable created by the difference between ConWX and RVM) and there is a lot of difference compared with the previous table.

Considering the sites of the best models, sites from Turkey were the ones that achieved lower R2. One of the R2 of Turkey's masts reduced its value into 0.58 which was low compared with the others. The next lowest value stood around 0.72 and was given by the other two masts of Turkey. However, as shown in *Figure 9* it was already known the difficulties Turkey would cause.

As mentioned before, none of the most adequate models had minimum BIAS, but they did with RMSE and MAE values. These errors had not the same pattern as the R2 and correlation. It occurred with the case of Turkey, that the error values were the highest and R2 and correlation lowest, but in the case of NORvar M32 in Norway, the errors were one of the highest (after Turkey), as well as the R2. Generally, the values obtained for BIAS were low, but sites with more data points had lower BIAS than others. WALbry M578 was the site with the lowest BIAS value with 4 years of data. This was followed by NIRnid M2 with 5.2 years and FRAotr M510 with 6.4 years.

6- and 10-fold cross-validation

In the case that the number of folds was increased to obtain better results, (this is in 6- and 10-fold cross-validation comparison for the best model) when comparing them between each other, when the data rows were larger (in those mentioned before, for example), 10-fold CV had better behaviour in errors and the difference in the values of RMSE was larger between two CVs. In the case of the R2 and correlation was also better in 10-fold but not in correlation, where 6-fold obtained higher results. Generally, 10-fold CV obtained less BIAS, RMSE, and MAE.

Comparison between 5-fold CV and 6- and 10-fold CV

Comparing these two with 5-fold CV, 6-fold CV was overlapped by 5-fold CV, since this was better in the metrics that 6-fold used to win to 10-fold. As mentioned, 10-fold CV obtained less BIAS, RMSE, and MAE, but 5-fold CV achieved to obtain higher R2 and correlation values. It should be highlighted that the only metric that made the difference between the others was the BIAS in 10-fold, between the other metrics the difference was not enough remarkable to assess one better than others. Even though, the effect of this CV would be assessed afterward in the section of Power Approach.

Another aspect to mention is that the error was not lower in those sites in which the number of k was a divisor. There was not found a pattern that followed this assumption.

All mentioned before is visualised in the following figures (*Figure 10*, *Figure 11* and *Figure 12*), all the sites are plotted, and each site is assigned to three bars. The first bar is assigned to 5-fold CV, the second to 6-fold and the third to 10-fold.

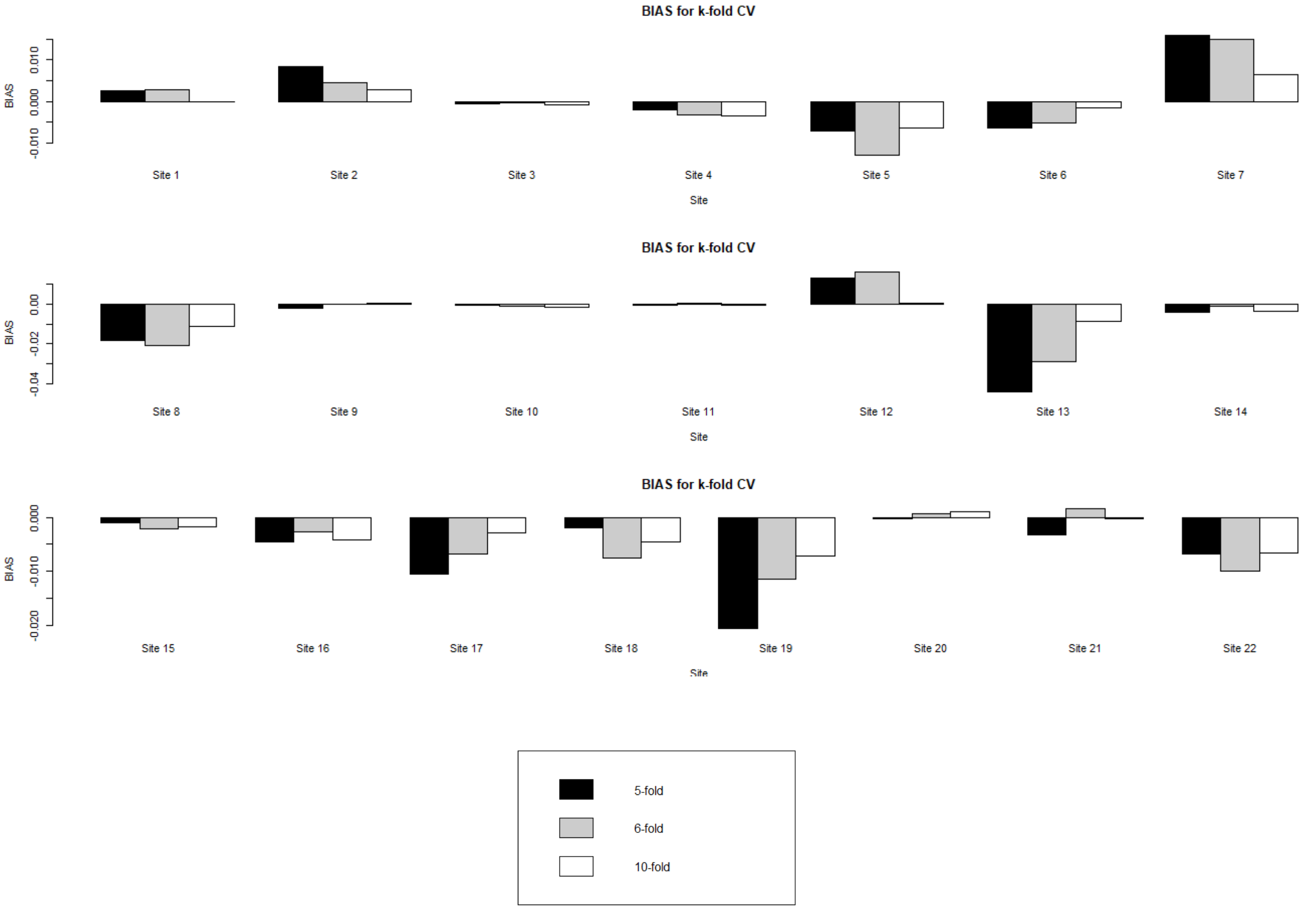


Figure 10 - Comparison of BIAS in 5-, 6- and 10- fold CV

In *Figure 10,* it is affirmed that 10-fold CV had principally lower BIAS than the others but a pattern was not found to assess this fact.

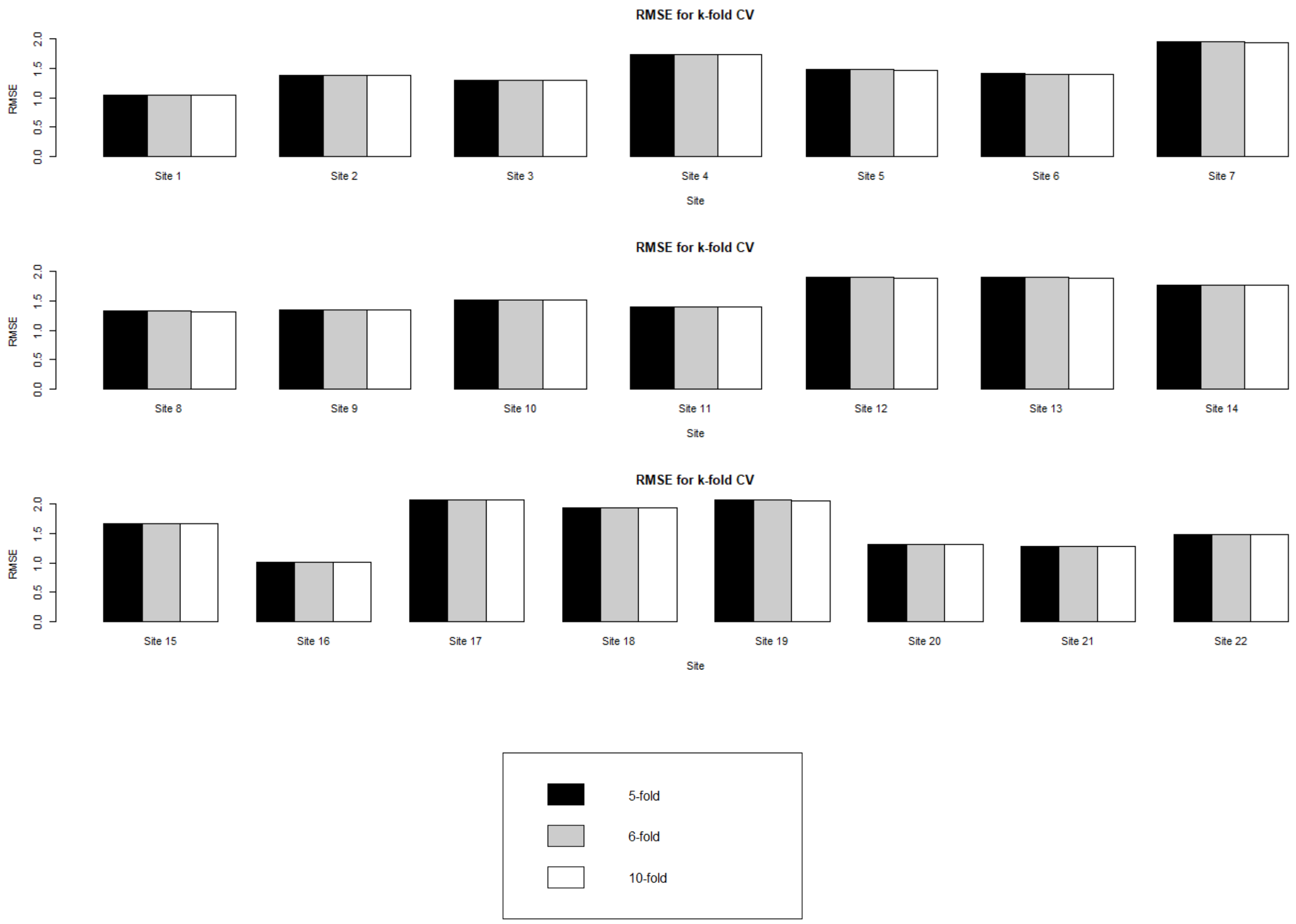


Figure 11 - Comparison of RMSE in 5-, 6- and 10- fold CV

In *Figure 11*, it can be concluded that globally 10-fold had again lower error than the other folds. MAE was not shown since it had the same pattern as RMSE. It should be highlighted that there were sites that are exceptions when following the pattern. Another aspect that it was concluded from here was that site that had a similar location and indeed similar data length used to have similar errors. This can be seen between Site 2 and 14 (North of England and South of Scotland), Sites 9,10 and 11 (North of Ireland), Sites 17,18 and 19 (Turkey) and Sites 20 and 21 (Wales).

This is not the case of the R2 plotted in *Figure 12*. Here, 5-fold CV is the CV that higher values obtained for this evaluation metric.

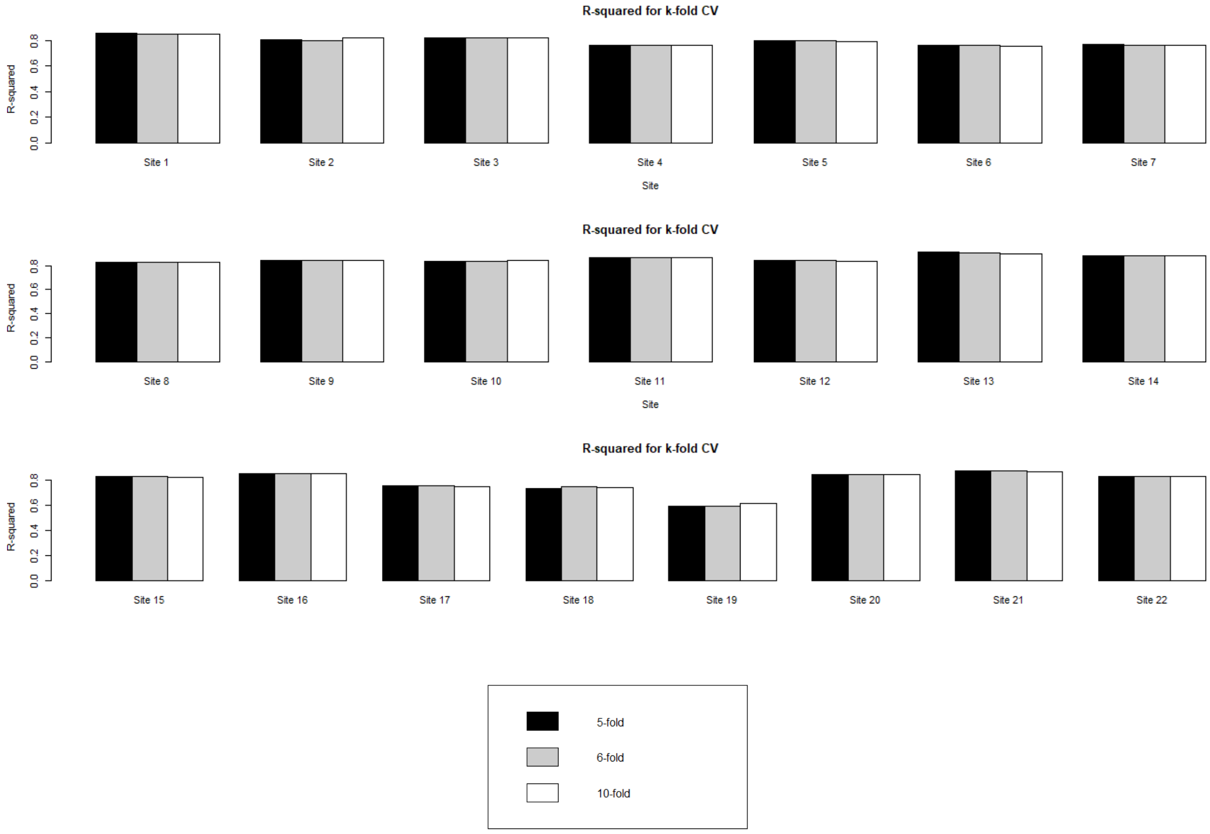


Figure 12 - Comparison of R2 in 5-, 6- and 10- fold CV

Mean of reanalysis tools

While testing the CV, a new column was included in the datasets with the mean of the four datasets available. It should be noted that every reanalysis tool should not gain the same weight in the model, and this new column took each of them as they would. Results are represented in *Appendix II*. Models discarded in the previous analysis were restored for this one and they obtained better results than in the previous analysis. However, in the findings, only the previous best models were used for the comparison.

Table 5 - Models used for the mean of the reanalysis tools in R language

- for1<- Site~ MRT + sinh + cosh
- for2 <- Site~ bs(MRT,3) + sinh + cosh
- for3 <- Site~ MRT + sinh4 + cosh4 + sinh + cosh
- for4 <- Site~ bs(MRT,3) + sinh4 + cosh4 + sinh + cosh
- for5 <- Site~ bs(MRT,3) + sinh4 + cosh4 + sinh + cosh + sinh6 + cosh6
- for6 <- Site~ bs(MRT,3) + sinh + cosh + sinh6 + cosh6
- for7 <- Site~ bs(MRT,3) + sinh4 + cosh4 + sinh6 + cosh6
- for8 <- Site~ bs(MRT,3) + sinh6 + cosh6
- for9 <- Site~ poly(MRT,2)
- for10 <- Site~ poly(MRT,2)+ sinh + cosh
- for11 <- Site~ poly(MRT,2)+ sinh4 + cosh4
- for12 <- Site~ MRT + sinh4 + cosh4
- for13 <- Site~ poly(MRT,2)+ sinh6 + cosh6

For this new variable, the model that most of the times achieved the lowest BIAS was the for12, a polynomial regression with 4π periodicity. RMSE and MAE coincided that the best model to obtain both lowest errors was the model 10 and this model was followed by models 5 and 9. The total highest R2 was obtained by model 5 and this once again achieved the best results for the best correlation, followed by formula 10.

As formula 5 was equivalent to the model that stood out in the previous analysis, this was taken for the implementation of the study. Formula 10 of this analysis was a polynomial model that was not specially highlighted earlier. The model that worst results obtained from this analysis was the model 12.

Even if the worst-case scenario was used, much better findings were achieved with this new variable. RMSE and MAE values were much lower in this analysis, reducing some of those to less than 1 and not surpassing the value 2 in RMSE. R2 values were higher than in the previous analysis. Even for the worst case that Turkey obtained in the previous section, here, R2 increased 0.2. The only evaluation metric that was higher in this analysis was the BIAS but having wee numbers did not make a difference. Moreover, the site's R2 is more balanced in this analysis. For the correlation, this was higher too in every site, not being lower than 0.88 which was provided by Turkey again.

Comparing best and worst cases of the mean, it should be highlighted that although it was mentioned before that the fifth model was not completely the best one it was chosen because it jutted out from the others and it was also taken for the previous analysis. For this reason, when analysing the worst model, it sometimes had better results than best cases, but the difference in the evaluation metric when model 12 was better was wee and wider when model 5 stood out.

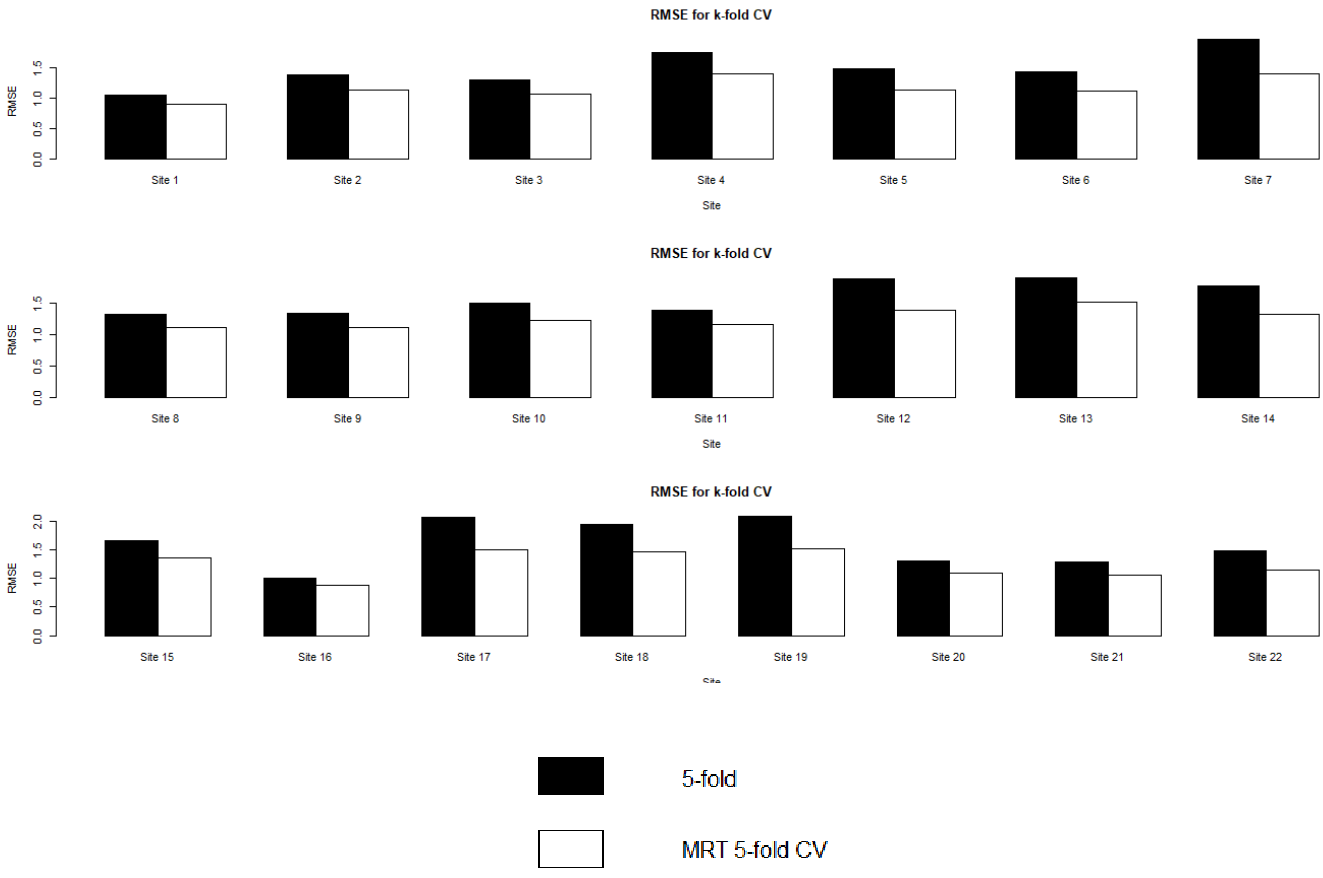


Figure 13 - Comparison of RMSE in 5- fold and MRT's best model

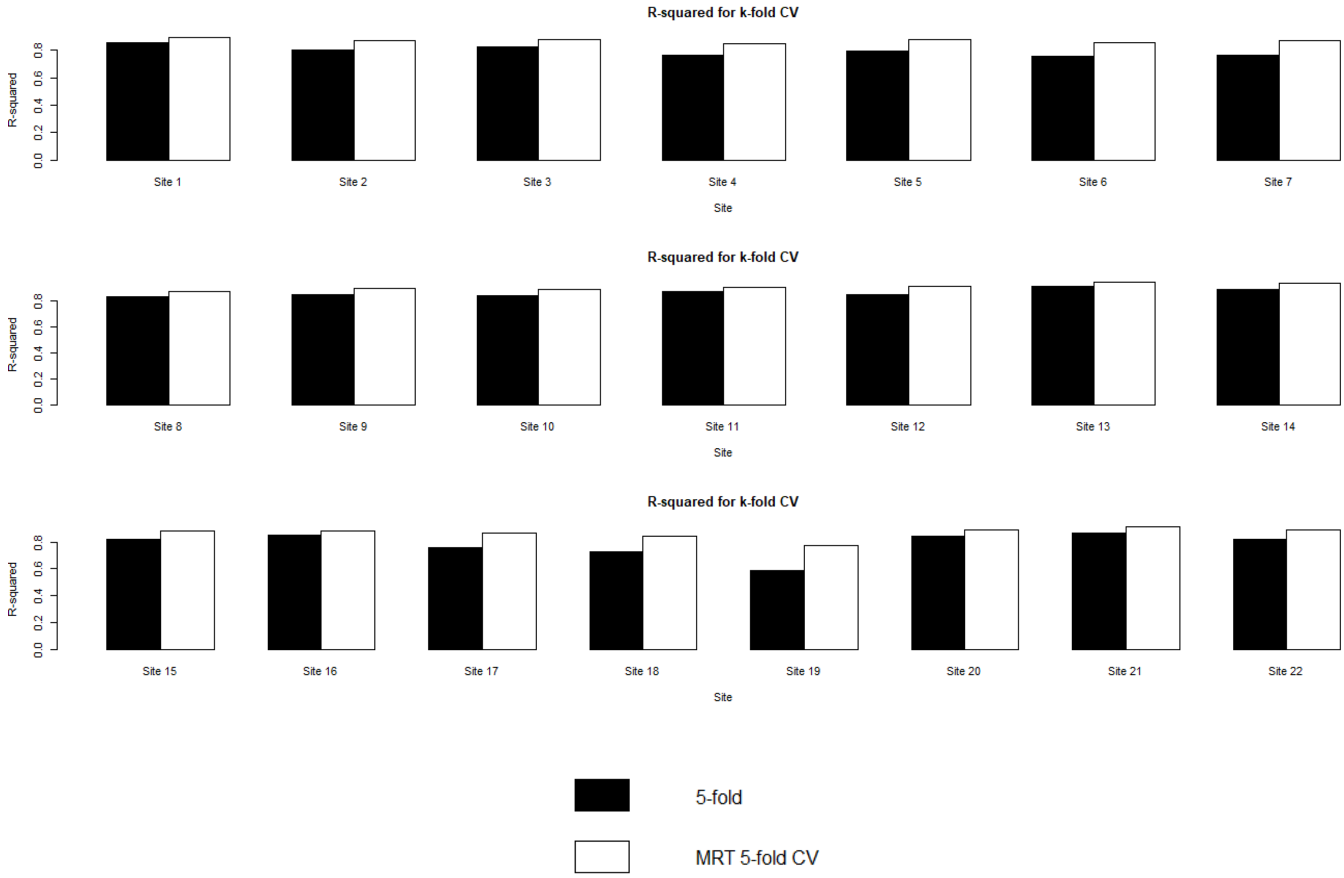


Figure 14 - Comparison of R2 in 5- fold and MRT's best model

Now, in *Figure 13* and *Figure 14*, RMSE and R2 were plotted again in a comparison between 5-fold CV and 5-fold MRT CV to observe how much better results were obtained with the application of the mean to the reanalysis tools. This time the first value of the site is assigned to 5-fold and the second to MRT values. Here, the pattern concluded before about the similarities between the sites from the same region and the similar data length was more consistent.

7.2. Backward Chaining Cross-Validation testing

Backtesting or backward CV is a random process which has different results in each iteration. Three iterations are represented in *Appendix II*. However, more iterations were done for the efficient performance of the model. It was assumed to get higher errors than in k-fold cross-validation, but those errors must reduce when the training length is larger. This CV made more difficult the decision-making of the model, but the total best correlation was obtained by model 36 which was used in both previous analysis. For this reason, in this CV the same model was used. It should be repeated and highlighted that as this CV is random, and although the best correlation always stood at model 36, the efficiency of the model changed in each iteration. The iterations that could contrast between each other are shown below.

For this first iteration, the randomness of the process seemed to affect its performance. Iterations that had better results were also obtained but the ones that most representatives were, are plotted.

Sites GERlir M1 (Site 6) and SWErod M39 (Site 16) were the sites with fewer training lengths, the first with 63 points and the second with 106, this was reflected in the errors which increased to 8 and 5. Values that were not assessed until now in the previous CV. However, site 8 had both, short training and testing lengths and it obtained a usual value of RMSE.

Correlation decreased importantly while training length window was short. For the first time, it was obtained a negative correlation and a value that was near 0 in SWErod M39, however, the error this site obtained was not such high, compared with the error GERlir M1 obtained. These values did not affect negatively to R2 but it did to BIAS which achieved a value of 2.

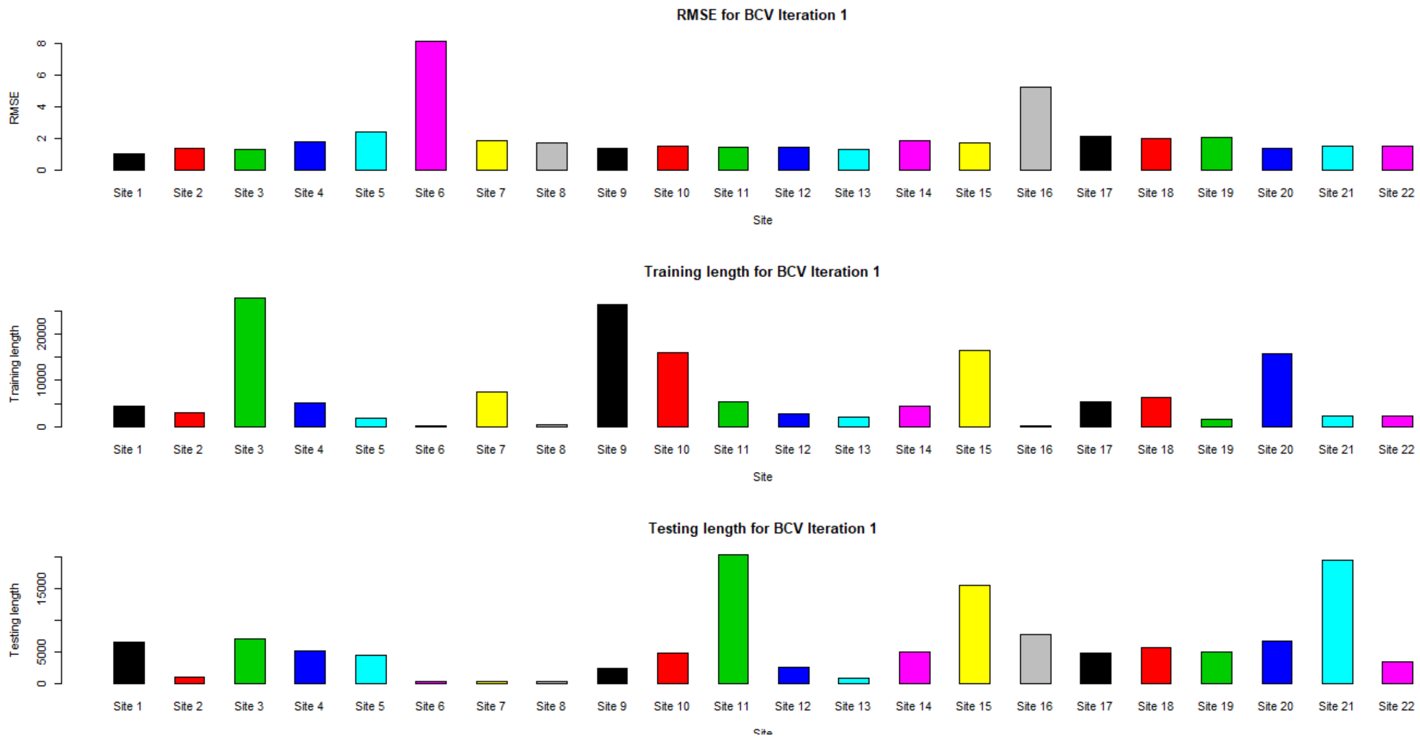


Figure 15 - Comparison of RMSE with training and testing lengths for each site in iteration 1

In *Figure 15*, RMSE is plotted in each site and compared with the training and testing length used in each site to affirmed what mentioned before. The highest errors were obtained where apart from the training length, also the testing window is narrow. For those in which the training length was wide and the testing length is short, the RMSE achieved acceptable values. Even though, short testing length added to short training length was the factor that most affect the value of the error.

In the second iteration, another iteration with short data lengths was chosen to analyse the effect this had and for the best understanding of the graphs. Here, Site 7 and 13 are the sites that worst results obtained and also the ones that least training length were assigned. However, in the last site, the error obtained was low and the training had similarities with site 13 in the training lengths, testing lengths and also in the total length of the data.

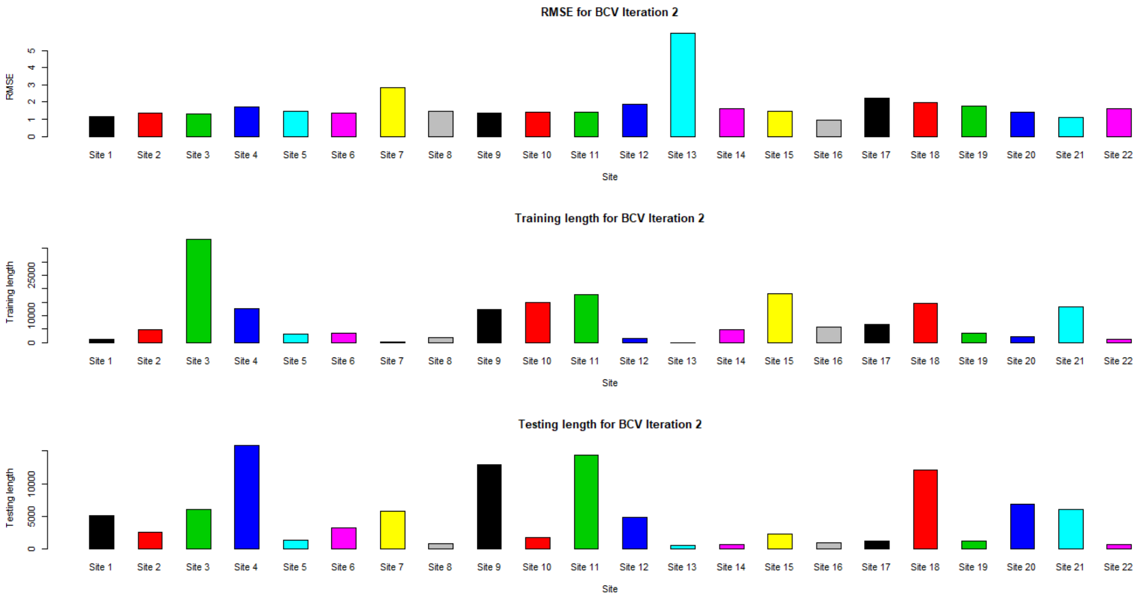


Figure 16 - Comparison of RMSE with training and testing lengths for each site in iteration 2

Finally, in the third iteration, the same pattern is repeated. Here, the shortest training length is assigned to GERlir M1 again, with 31 points this time. This time the testing length is much longer, but the RMSE error stood for 71.71. The BIAS and MAE errors were high compared with other values from different sites and despite having a high R2 (probably because of the long test length), correlation represented a value near to 0.

There are not a lot of short test lengths in the iterations, but in those that its value is between 0 and 200, the R2 reduces slightly its value to 0.7.

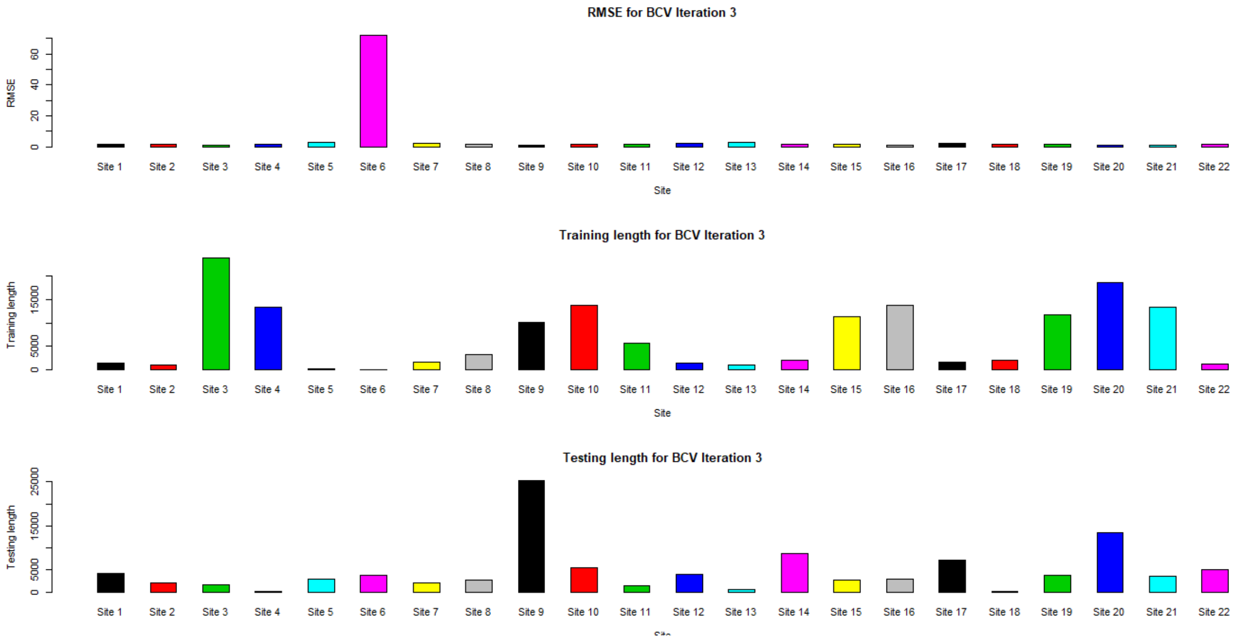


Figure 17 - Comparison of RMSE with training and testing lengths for each site in iteration 3

In *Figure 17*, the fact that the value of an RMSE was 71, makes the other errors negligible and this last graph is not as visual as the previous ones.

Generally, as expected, errors were higher and model efficiency lower, but this CV also can provide good results. Other comparisons between correlation, R2, and training lengths are represented in *Appendix III*.

So, having observed *Figure 15*, it is interesting to look at the graphs plotted below (*Figure 18*). While the training length was increasing, it was common that the RMSE values also did.

In Site 6 the graph does not show the value of 71 since the rest of the RMSE for different sites seem negligible. But all sites do not seem to follow this pattern since if it is observed in *Figure 18 d)* some of the errors increased with the training length. It cannot say that it follows the same pattern with testing length, because testing length is dependent on the training length when assessing the RMSE.

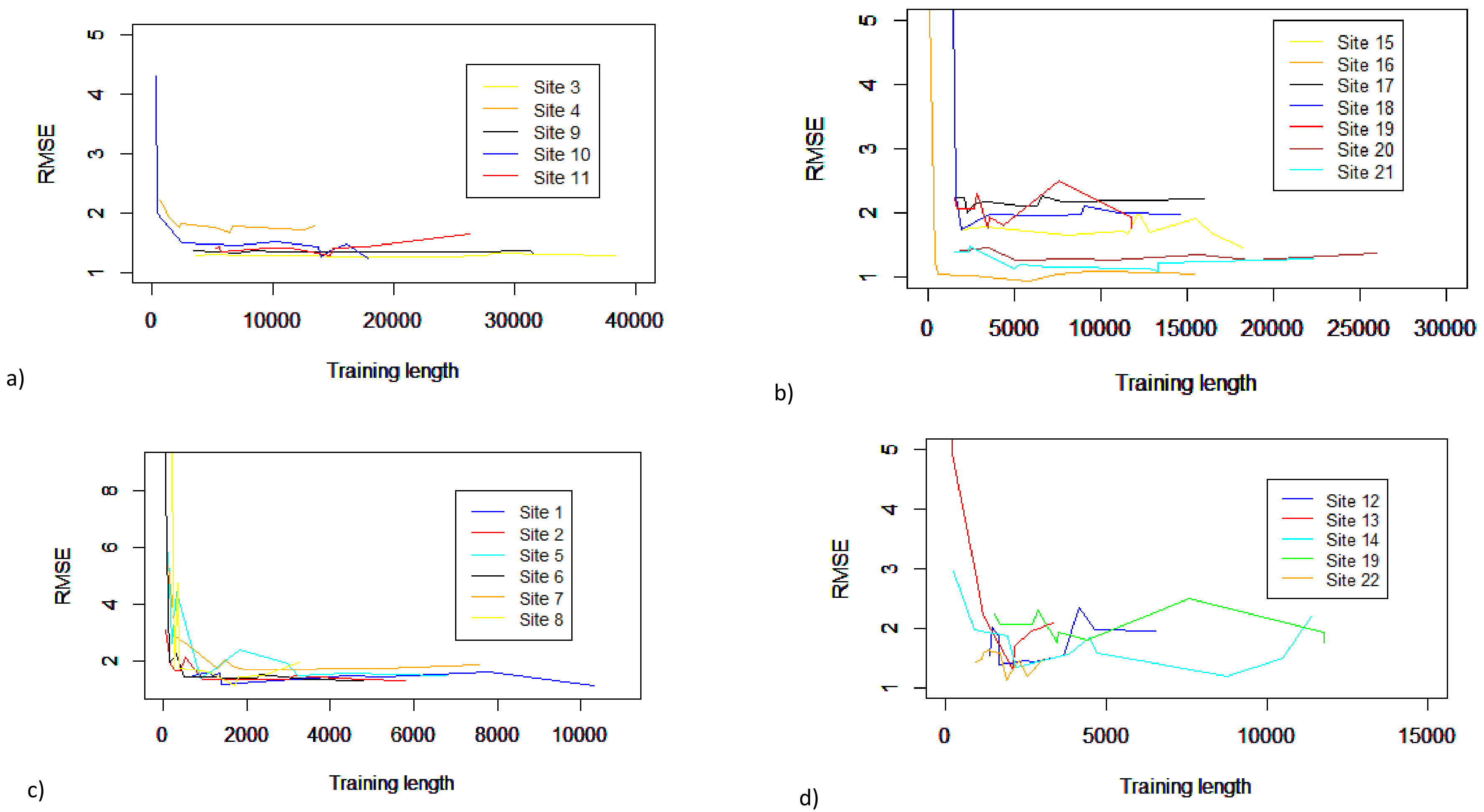


Figure 18 - RMSE vs Training length in BCV

8. Power Approach

For each cross-validation and each iteration, predictions were saved for the approach of the power. For this, the predictions were compared with the power curve of different turbines. The turbines were chosen depending on their hub height and compared with a similar hub height mast. After this, the power density of some sites is represented.

For this analysis, it should be highlighted that wind speed was not the only variable to consider when assessing the power. Density, which changes with the site and height and different turbines were relevant variables to be considered, as mentioned in Uncertainties section.

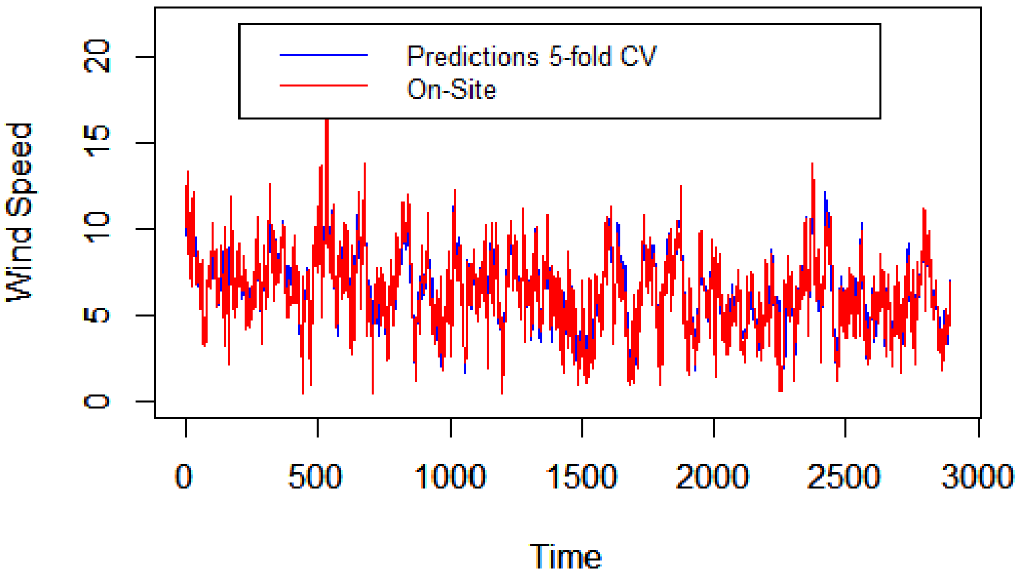
8.1. Enercon E-40/500

The first turbine used is the Enercon E-40/500, its characteristics are taken from (Enercon, 2019). From this, the rated power, the hub height (50.3 m), and the power curve were taken and considered for the assessment. The sites compared with the power curve of this turbine were Sites 1, 11, 14 and 17. Most of the sites managed to obtain the rated power in all the sites, besides the first one, where it is close to it in every CV. It is concluded that as the site's data is not wide, 10-fold iteration takes tiny training/testing lengths and affect the power output.

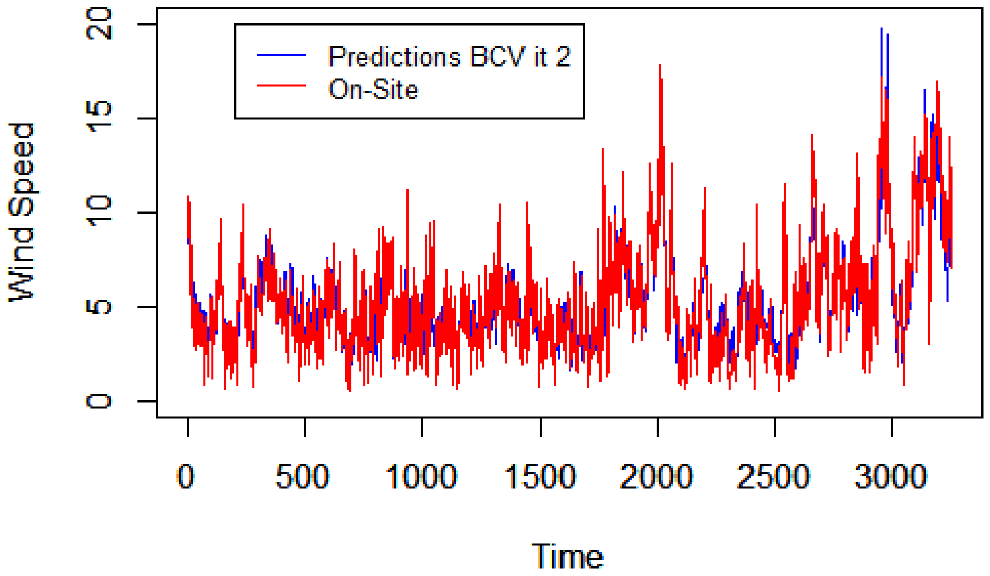
Table 6 - Maximum power obtained from ~ 50.3 m mast heights

Mast	*Height (Hub height 50.3 m)*	*k-5 fold*	*k-6 fold*	*k-10 fold*	*Backward CV Iteration 2*
1 ENGgra	*60.13*	*498.5673*	*498.5469*	*449.9734*	*500*
11 NIRnid	*50*	*500*	*500*	*500*	*500*
14 SCOdun	*50.2*	*500*	*500*	*500*	*500*
17 TURcig	*63.24*	*500*	*500*	*500*	*500*

In *Figure 19* a comparison between on-site data and 5-fold CV was done first. Predictions values were lower than on-site data ones, so they did not reach the value of 15 m/s, so neither the rated power. This was not the case in *Figure 19 b)* where the predictions of BCV achieved the maximum power.



a)



b)

Figure 19 - Comparison between On-site data and a)5-fold CV and b) BCV iteration 2 predictions made for Site 1, 50 m masts

From the previous figure, it is concluded that k-fold cross-validations obtain more constant values and it did not have as many spikes as BCV.

Once the predictions of wind speed were analysed, it was observed how they affected the power density in *Figure 20*. The fact that BCV had larger testing length made the predictions obtained the higher and the lower power values. The spikes mentioned before, made BCV obtain the rated power.

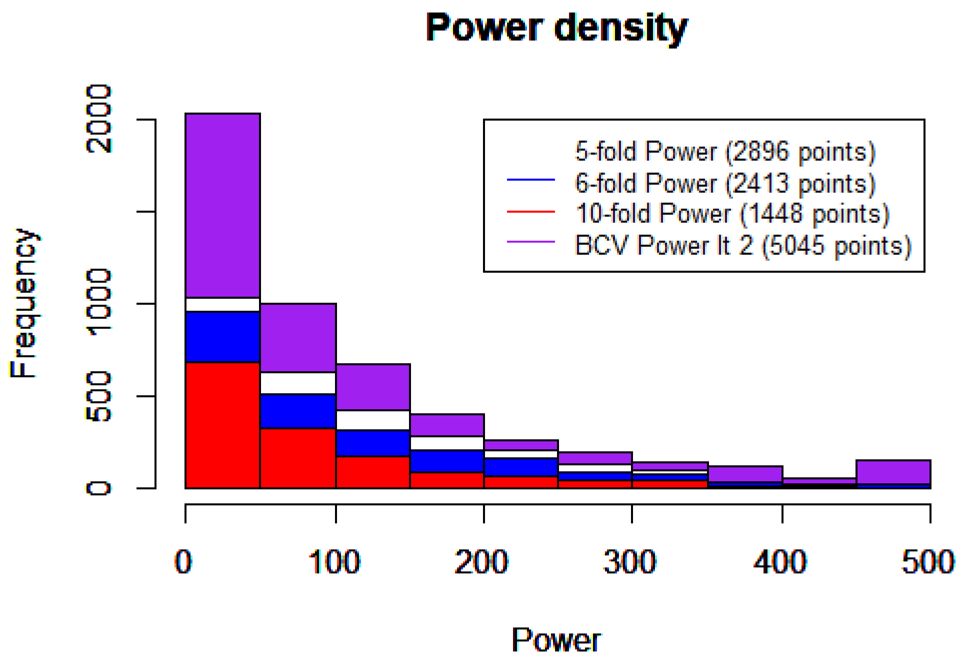


Figure 20 - Power density for Site 1

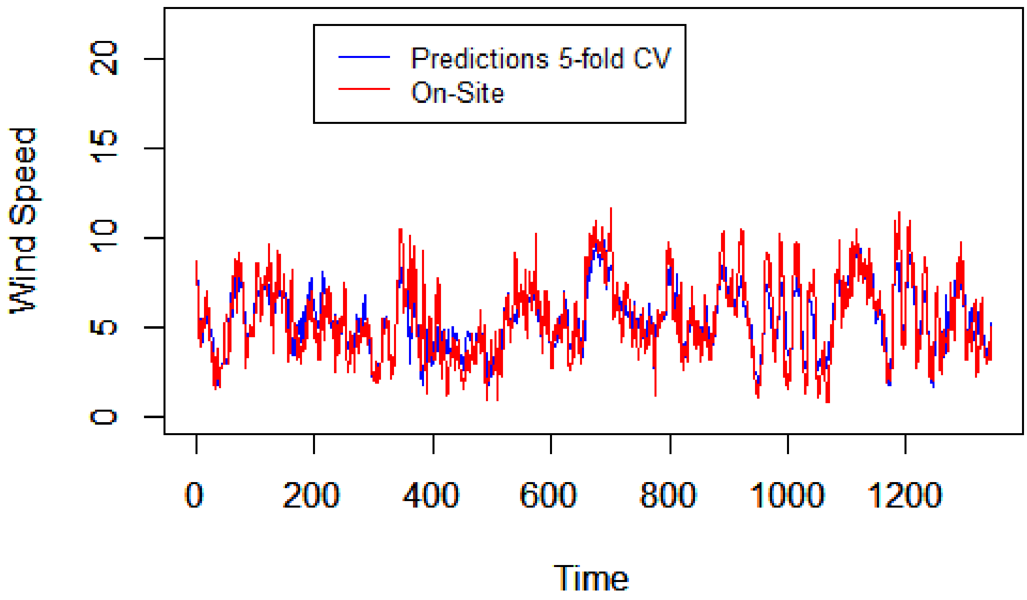
8.2. Siemens SWT-2.3-113

For the masts that were 100 m high, a Siemens SWT-2.3-113 was chosen and its characteristics can be found in *(Wind Turbine Models, 2019).* Here, the sites with similar hub height are sites 3, 5, 6, 7 and 8.

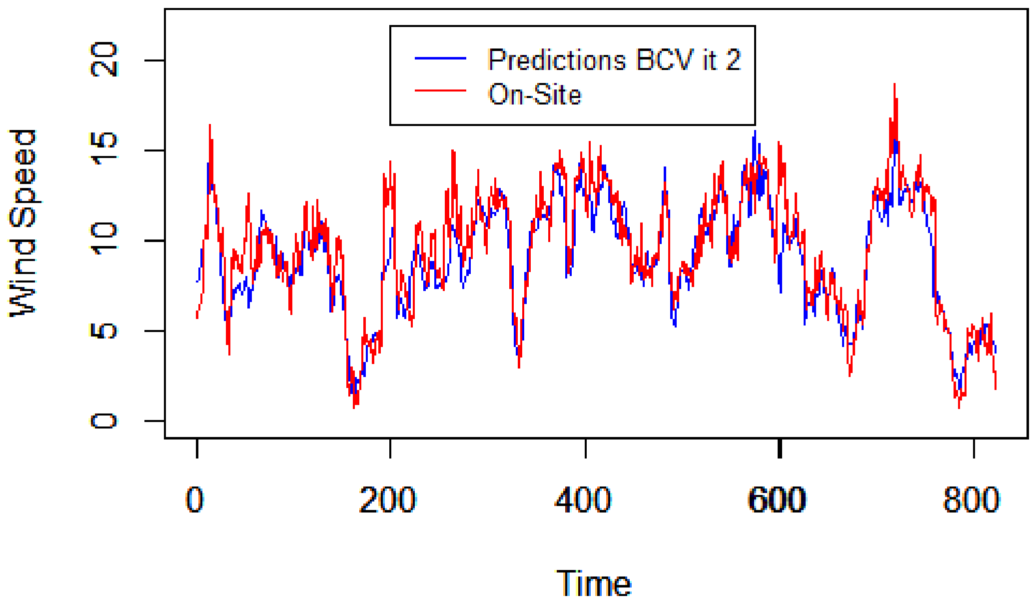
Table 7-Maximum power obtained from ~ 99.5 m mast heights

Mast	*Height (Hub height 99.5 m)*	*k-5 fold*	*k-6 fold*	*k-10 fold*	*Backward CV Iteration 2*
3 FRAotr	*102*	*2300*	*2300*	*2300*	*2300*
5GERlac	*100.73*	*2300*	*2300*	*2300*	*2300*
6 GERlir	*100.73*	*2300*	*2297.127*	*2297.232*	*2300*
7 GERpr1	*100*	*2300*	*2300*	*2300*	*2300*
8 GERpru	*118.5*	*2019.165*	*2025.769*	*2028.682*	*1830.841*

Same occurs in *Table 7*, where sites 6 and 8 did not manage to obtain the rated power. GERlir in 6- and 10-fold and GERpru in all of them. This may be concluded for having the shortest data points (and training lengths).



a)



b)

Figure 21- Comparison between On-site data and a)5-fold CV and b) BCV iteration 2 predictions made for Site 8, 100 m masts

In *Figure 21*, the same pattern can be concluded. BCV had several spikes that make more irregular predictions which are represented afterwards in *Figure 22*. 5-fold CV did not even obtain 13 m/s value so neither the rated power. This can be easily seen in the following figure, where 5-fold CV having more testing points than BCV, it took the lowest values of the power.

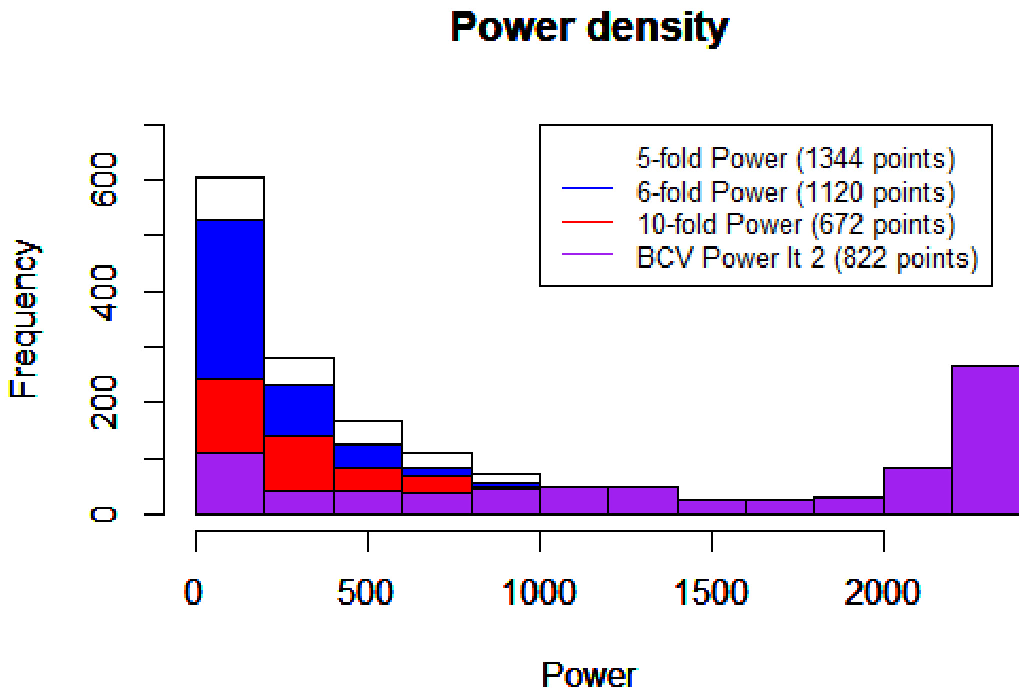


Figure 22 - Power density for Site 8

8.3. GE Energy 2.75-103

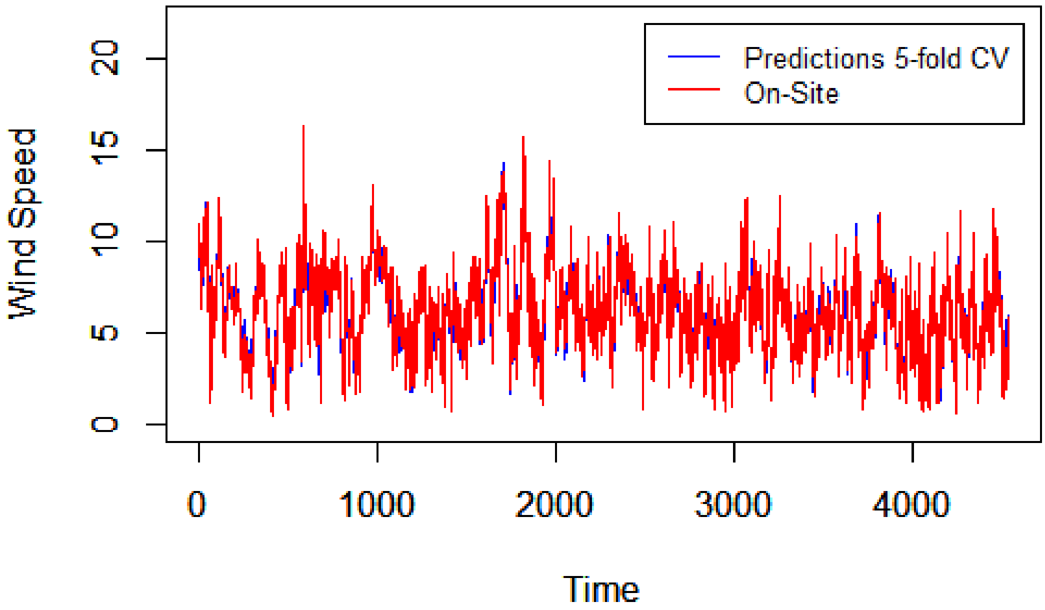
For the masts that are 75 m high, GE Energy 2.75-103 was chosen, and its characteristics can be found in *(GE-Energy, 2019)*. Finally, sites left were assessed.

Table 8 - Maximum power obtained from ~ 75 m mast heights

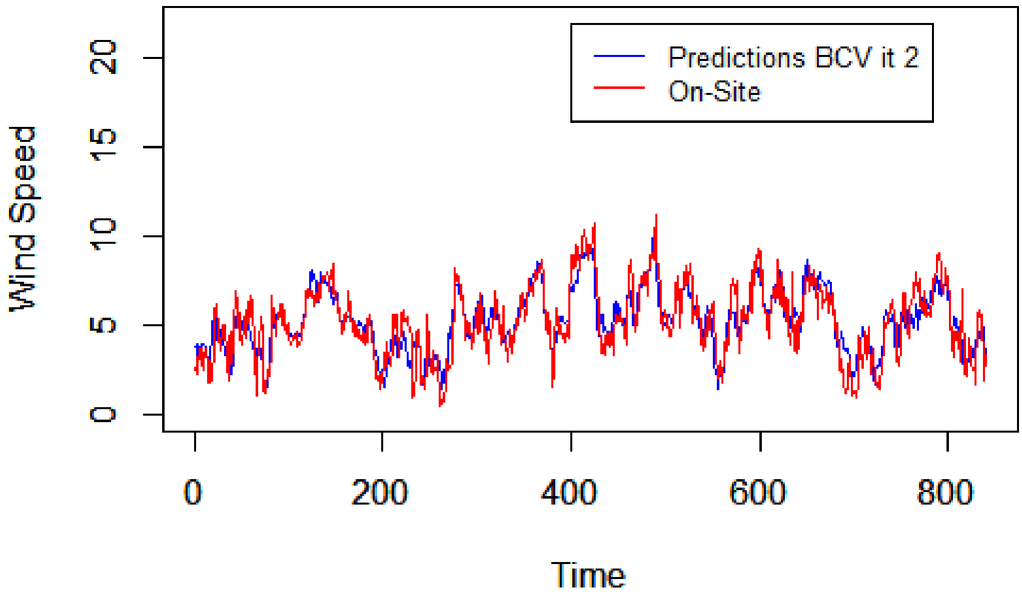
Mast	*Height Hub height (75 m)*	*k-5 fold*	*k-6 fold*	*k-10 fold*	*Backward CV Iteration 2*
2 ENGpkh	*80.1*	*2750*	*2750*	*2740.447*	*2750*
4 FRApdl	*78.05*	*2750*	*2750*	*2750*	*2750*
9 NIRccg	*80.31*	*2750*	*2750*	*2750*	*2750*
10 NIRcgr	*80.44*	*2750*	*2750*	*2750*	*2750*
12 NORskv	*83.9*	*2750*	*2750*	*2750*	*2750*
13 NORvar	*83.66*	*2750*	*2750*	*2750*	*2750*
15 SCOfre	*75*	*2750*	*2750*	*2750*	*2750*
16 SWErod	*79.99*	*2750*	*2750*	*2589.993*	*2724.28*
18TURevr	*81*	*2750*	*2750*	*2750*	*2750*
19 TURhvz	*81*	*2750*	*2750*	*2750*	*2670.364*
20 WALbry	*79.59*	*2750*	*2750*	*2750*	*2750*
21 WALglh	*70.1*	*2750*	*2750*	*2750*	*2750*
22 WALmmr	*70.6*	*2750*	*2750*	*2750*	*2750*

More sites were added to this analysis group since their mast heights were between the other two turbines' hub heights.

Here, all the 5- and 6-fold predictions achieved to obtain the rated power. There were weaknesses on the other two CV predictions in site 16, and with backward CV in site 19. Following the same analysis of the previous turbine, the testing length for site 19 was more limited (106 points) than in any other k-fold it happened in the months in which the wind is gentle, thus, the power was again lower. In the case of site 16, the testing length of backward was larger than the others, before, even though this CV obtained almost the rated power, and 10-fold did not. This time, the length of the datasets was not as short as in the previous analysis.



a)



b)

Figure 23 - Comparison between On-site data and a) 5-fold CV and b) BCV iteration 2 predictions made for Site 16, 75 m masts

Finally, in *Figure 23*, the pattern mentioned before was weak due to the testing length of the BCV and 5-fold cross-validation seemed to have more spikes. In *Figure 24* the power density of this is provided again. Testing points are so different that it is unavoidable to k-fold cross-validation to obtain better results. Nevertheless, 10-fold is still weak in the highest values of power.

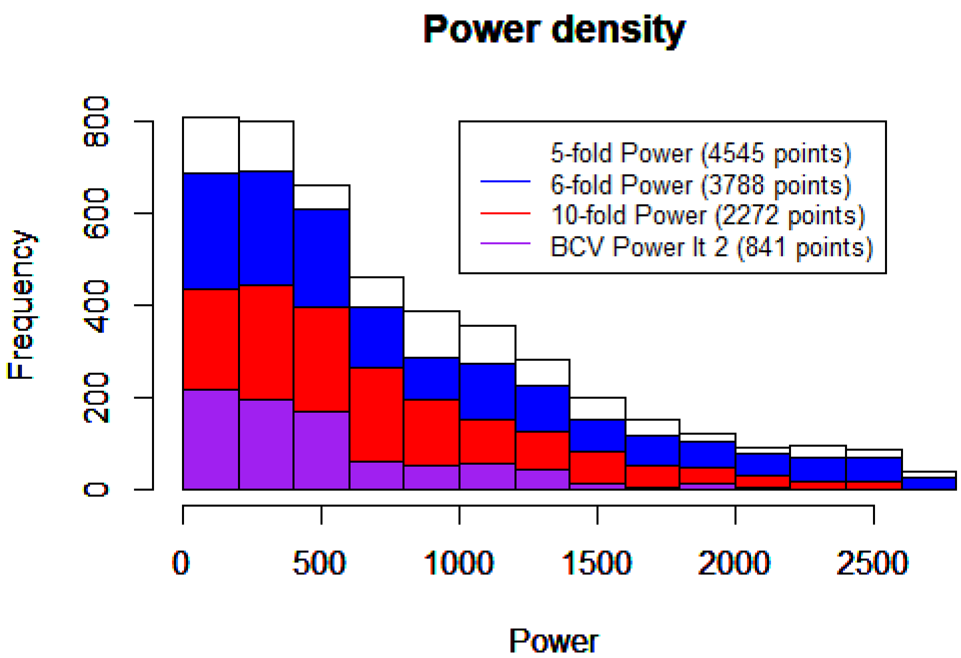


Figure 24 - Power density for Site 16

9. Discussion

Following the document structure, the first discussion point is given to the individual analysis of the reanalysis tools. Here, MERRA was the tool that larger distances had to the masts, however, there was not found a pattern that could justify its efficiency in the linear model. It could be said that it could improve its performance in sites that have a similar height to the measurement height of MERRA, but it did not match with the distances to the mast. Nevertheless, its efficiency is not the worst. For the others, a pattern was found.

The observation of k-fold cross-validation between RMSE errors and training and the testing length was not useful since error between k-folds assessed were very similar and many folds must be added to increase the error to a high value.

It is recommended that the best k to fit a model to the database is the one that is a divisor of the total data length. In this case, when k=5, the sites that are multiples to that were the sites 6, 8, 14, 20 and 21, when k=6, sites 8, 12 and 16, and when k=10, sites 6, 8, 14, 20. Considering this, the results analysed did not seem to include a pattern that matched with that.

MRT, being the mean of the four-reanalysis data, it assigned the same weight to all the reanalysis tools, but these were not equally good when assessing the wind speed. All in all, all the reanalysis data demonstrated a similar tendency since the wind had a daily, monthly and a yearly pattern. This can be termed as multicollinearity, and it tends to produce larger errors. In *Figure 25*, the coefficients of the multiple linear models are represented, where they seem to behave adequately since all of them were positive and it was similar for every dataset, so this option was discarded.

The contribution that each reanalysis tool has in the linear model would depend on the height of the mast, the height of the reanalysis tool, the efficiency of the reanalysis tool and the distance this has to the mast.

```
Call:
lm(formula = Site ~ ERA5 + MERRA + ConWX + RVM, data = lists[[1]])

Coefficients:
(Intercept)         ERA5        MERRA        ConWX          RVM
     0.8224       0.4759       0.1378       0.1913       0.0574
```

Figure 25 - Linear model coefficients for ENGgra site

In *Figure 25,* ERA5 is the reanalysis tool that most contribution had in the model. However, it should be considered that the mast was located at 50 m, as well as MERRA and ConWX and having a distance to it of 24.9 and 1.1 km. ERA5 grid point was located at 100 m and 10.1 km far from the mast. This happened in lots of the sites analysed. This discussion point is also linked to the first discussion point of this section.

Also, for the identification and avoidance of multicollinearity, correlation matrixes (*Figure 26)* were analysed. Here, Turkey case was the worst with a -0.24, a value which discarded the collinearity issue.

```
        Site  ERA5 MERRA ConWX   RVM  sinh  cosh sinh4 cosh4 sinh6 cosh6
Site    1.00  0.75  0.73  0.82  0.80 -0.12 -0.05  0.01  0.03  0.00 -0.01
ERA5    0.75  1.00  0.77  0.78  0.68 -0.24 -0.14  0.01  0.00  0.03 -0.02
MERRA   0.73  0.77  1.00  0.83  0.71 -0.03  0.01 -0.03  0.01  0.01 -0.02
ConWX   0.82  0.78  0.83  1.00  0.82 -0.17 -0.03  0.00 -0.03  0.01 -0.02
RVM     0.80  0.68  0.71  0.82  1.00 -0.14 -0.02  0.04  0.03  0.01 -0.01
sinh   -0.12 -0.24 -0.03 -0.17 -0.14  1.00  0.00  0.00  0.00  0.00  0.00
cosh   -0.05 -0.14  0.01 -0.03 -0.02  0.00  1.00  0.00  0.00  0.00  0.00
sinh4   0.01  0.01 -0.03  0.00  0.04  0.00  0.00  1.00  0.00  0.00  0.00
cosh4   0.03  0.00  0.01 -0.03  0.03  0.00  0.00  0.00  1.00  0.00  0.00
sinh6   0.00  0.03  0.01  0.01  0.01  0.00  0.00  0.00  0.00  1.00  0.00
cosh6  -0.01 -0.02 -0.02 -0.02 -0.01  0.00  0.00  0.00  0.00  0.00  1.00
```

Figure 26 - Correlation matrix of Site 17 for the identification of collinearity between variables

However, a solution to apply for the unknown reason would be the gradient boosting machines, which is an automatic learning technique used for regression analysis and for statistical classification problems. In order to reduce the BIAS and the variance, it produces a predictive model converting weak learners into strong learners. It is like other types of regression in which they consider other independent variables to explain a final variable. As a tandem method for regression, it uses decision trees optimization, in which its model is a tree and its consequences. The objective of this is to clarify in a determined level of the tree which variables must be considered *(Mark Landry, 2016).*

As it was known, backward CV had larger errors than a k-fold CV. Even though they were expected larger, they only achieved excessive values for wee training lengths, generally, lower than 200-point lengths. The randomness of this process did not result negatively for the results obtained. However, the fact that the testing window is followed by the training window, made that if the windows are narrow, a lot of months of the year not being assessed, the model could be only trained and tested in the same season, or by contrast in different seasons, providing better or worse results based on that.

Whereas it has been clarified that the error was reduced while the training length increased. In the case of the testing length, this is dependent on the training length. It needed a short window of training length and a short window of testing length to have a high value of the error. This was not perceived in the other way around.

Moreover, backward CV got some negative results for wind speed, which cannot be possible in a real situation. However, as the linear model was set without any constraints, the values of wind speed that were near to 0 could lead to giving negative predictions. This can be easily perceived in *Figure 27*.

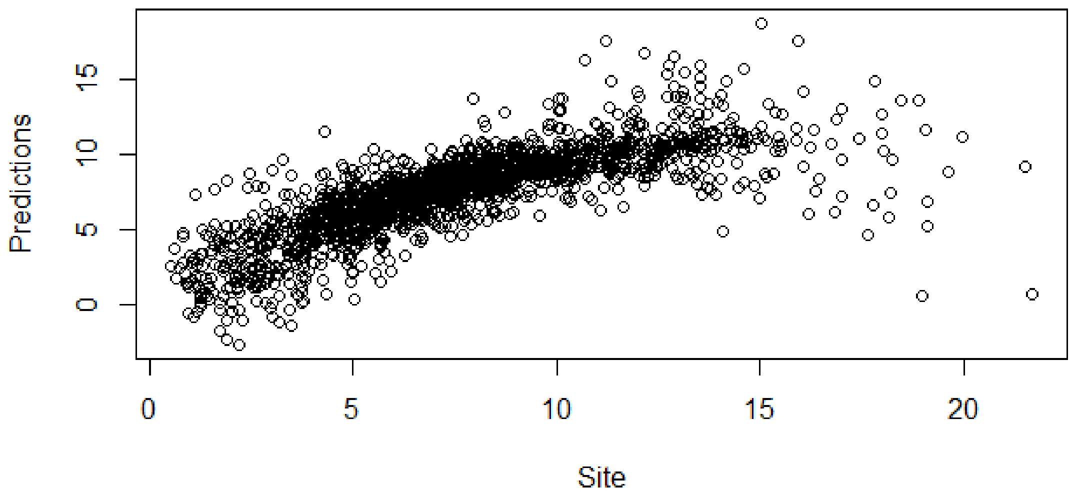


Figure 27 - Negative wind speed values

It is assumed that the wind speed has the range of being positive and a maximum of 25 m/s, although between 0-3 m/s there is no power production. However, data may fall outside the range. There are two possibilities to limit this, censoring the data or truncating it. The first would report that data at the range limit, the second omits it completely (Messner, et al., 2016).

There are some patterns to consider when assessing the power since it was linked to the training and testing lengths. This was harsher to analyse when those lengths were changing. First, although when results were analysed, there was not a k-fold number that was considerably better than the other, by contrast, it was in this case, where 5- and 6-fold dominated most of the times. 10-fold CV's predictions were weak for some sites, which means that so many folds did not allow the datasheet to train and test. In those cases where BCV was slacking off, the training and testing lengths were happening in short lengths and gentle wind seasons. It was detrimental for BCV to have a short test length.

From the predictions of the k-fold CV and BCV, it was concluded that BCV had much more spikes with which it allowed achieving the rated power. These can be convenient to assess when speaking about wind speed/power since it is an issue that wind power already has and being able to handle these spikes would be an advance in the development of the energy.

In terms of analysing the density of the power, BCV obtained better results, since it allowed to the turbine working in the rated power more times than the others (due to the spikes) and it was not severely affected by the data lengths used for the testing.

Much more observations could be completed with every site and every iteration but too many data are available. Variables to consider and are missed in this document could be aggregated for future works.

10. Conclusion

In the event of selecting an individual reanalysis tool, ERA5 is the reanalysis tool with the best efficiency in every site and model. This is followed by ConWX, which is the only reanalysis tool that can overcome ERA5's efficiency when all the factors are favourable. Same occurs with RVM, but this was not able to achieve better efficiency than ERA5. Finally, MERRA can achieve good contribution but there is no pattern that can define that.

This document uses a linear model to predict wind speed based on historical data. The simple linear model has been used in other studies for the wind speed resource assessment. It has been found that the linear model has a significant role in the field and can successfully be used for its measurement.

Being wind speed the only variable available, new variables were added to the linear model and within them, the most effective option was (in R language),

```
Site~ bs(ERA5,3)+bs(MERRA,3)+bs(ConWX,3)+bs(RVM,3) + sinh4 + cosh4 + sinh + cosh +
sinh6 + cosh6
```

which is composed of a cubic spline and three periodicities, 2 π, 4 π, and 6 π. This is the model that most times achieve the best performance (lower errors and higher R2 and correlation), even if it did not have a remarkable advantage from others.

In the k-fold cross-validation method, it was concluded that for the RMSE error (and MAE) that sites that were close to each other and their data lengths were similar, have similar errors between them. Between 5-, 6- and 10-fold CV, there is not a considerable change between the evaluation metrics. 10-fold is better for assessing RMSE and MAE but primarily the BIAS. 5-fold CV achieves better results for R2 and correlation. Moreover, a pattern was found in the error between datasets that are located next to each other in the map and indeed had similar data lengths.

Mean of Reanalysis Tool variable achieved better results than any other reanalysis tool. This was concluded to be since variables in the model selected had similarities, but it cannot be termed as multicollinearity since the correlation matrix of every site was analysed. Gradient boosting machine was chosen as an alternative option for future works.

Regarding the Backward cross-validation, it is obvious that its performance was based on the training and testing lengths. There was a limit of data points when assessing wind speed, and the larger the training/testing length (especially training length), the lowest the error and the higher the efficiency. By contrast, the maximum power obtained was affected by the testing length. However, there was a slight mishap with this cross-validation, as the predictions obtained had negative values. For this, the linear model restrictions were needed, such as censored or truncated regressions. This makes less robust the model that seemed to be strong.

Finally, predictions were evaluated with the power maximum and power density of the sites. 5-fold cross-validation's predictions were the predictions that most of the times achieve rated power, but there was a conflict between this and BCV in the power density. In the cases were BCV have short training lengths, the power density was more equilibrated (due to spikes) and higher values were obtained.

Both CVs proved that there are successful options when assessing wind resource, however, if BCV is accurate enough on assessing the spikes that wind speed produces in the generation of wind power, this cross-validation can be the best option for the long-term predictions.

All in all, based on the results and if some restrictions and improvements mentioned (and highlighted in future works in the following section) are applied to the model, these data analysis methods could replace erecting a met mast at a site of interest for development and ERA5 and ConWX reanalysis tools would be the best options in the case of not using the four of them.

11. Future Work

There was a first idea of the testing of the model in an earlier time period, but not having easy access to that and not even having mast data to compare it, it was pointless to do it. However, this must be added as future work to assess the efficiency of the cross-validations completely.

There are too many variables to consider when analysing a linear model, and other variables such as wind direction, temperature and humidity would help to make the study more accurate. It would be interesting to apply this model adding those variables. Moreover, there are other types of regression that could also be considered, such as penalized regression.

As negative values of wind speed were obtained in BCV, the linear model should be truncated for the most accurate performance of the model.

For the reduced errors of the model, a gradient boosting machine method could be implemented as mentioned in the section Discussion. It may be interesting to focus on reducing the error for the assessment of the spikes in the predictions.

Having analysed backward cross-validation, apart from limiting the linear model not to obtain negative results, this cross-validation could obtain better predictions with some other restriction that would be interesting to investigate in the future.

Finally, as the project objective was to conclude the reduction in the cost that would have the avoidance of the installation of a mast, a cost analysis should be done. The model selected was formed by the four-reanalysis tool and it should be considered if this would have savings in the resource assessment.

12. References

- Alencar, D. B. d., Affonso, C. d. M., Oliveira, R. C. L. d. & Filho, J. C. R., 2017. *Different Models for Forecasting Wind Power Generation: Case Study.* [Online]

  Available at:

  https://www.researchgate.net/publication/321390529_Different_Models_for_Forecasting_Wind_Power_Generation_Case_Study

  [Accessed 01 06 2019].
- Ammonit, 2019. *Accurate wind measurement with traceable wind data.* [Online]

  Available at:

  https://www.ammonit.com/en/wind-solar-messsysteme/windmessung

  [Accessed 01 06 2019].
- Ammonit, 2019. *THE IMPORTANCE OF WIND MEASUREMENT.* [Online]

  Available at: https://www.ammonit.com/en/wind/wind-measurement

  [Accessed 04 06 2019].
- Brownlee, J., 2018. *Machine Learning Mastery.* [Online]

  Available at: https://machinelearningmastery.com/k-fold-cross-validation/

  [Accessed 30 07 2019].
- C. Morton, A. G. J. K., n.d. *An Investigation into the Specification of an Off-Grid Hybrid Wind/Solar.* [Online]

  Available at:

  http://www.esru.strath.ac.uk/Programs/Merit/case_study_Ecobarn.pdf

- Cast Massey University, 2017. *Summarising Relationships.* [Online]

  Available at:

  http://cast.massey.ac.nz/core/bk/climatic_c3/text/Chapter_s_2.html?isSummar y

  [Accessed 2019 08 12].

- Cohn, J. P., 2008. How Ecofriendly Are Wind Farms?. *Bioscience,* 58(7), p. 576–578.

- ConWX, 2019. *ConWX.* [Online]

  Available at: http://conwx.com/services/power-forecasts/numerical-weather-prediction-data/

  [Accessed 28 07 2019].

- Dean Laslett, P. J. J., 2016. A simple hourly wind power simulation for the South-West region of Western Australia using MERRA DATA. *Renewable Energy,* 10.

- Deign, J., 2019. *Global Wind Market to Add 65GW per Year to 2027 - greentechmedia.* [Online]

  Available at: https://www.greentechmedia.com/articles/read/global-wind-market#gs.fvccyx

  [Accessed 31 05 2019].

- Delgado, R., 2018. *RPubs.* [Online]

  Available at: https://rpubs.com/rdelgado/405322

  [Accessed 30 07 2019].

- EMHART, M. P., 2018. *Exploring Machine Learning Models.* [Online]

  Available at:

https://upcommons.upc.edu/bitstream/handle/2117/117771/131678.pdf
[Accessed 2018].

- Enercon, 2019. *windpower.* [Online]
  Available at: https://www.thewindpower.net/turbine_es_164_enercon_e40-500.php
  [Accessed 13 08 2019].
- Energy.gov, 2019. *Energy.gov.* [Online]
  Available at: https://www.energy.gov/maps/how-does-wind-turbine-work
- Frost, J., 2018. *How To Interpret R-squared in Regression Analysis.* [Online]
  Available at: https://statisticsbyjim.com/regression/interpret-r-squared-regression/
  [Accessed 2019].
- GE Renewable Energy, 2018. *Wind Energy: How It Works.* [Online]
  Available at: https://www.ge.com/renewableenergy/wind-energy/technology/what-is-wind-energy
  [Accessed 01 06 2019].
- GE-Energy, 2019. *The Wind Power.* [Online]
  Available at: https://www.thewindpower.net/turbine_en_648_ge-energy_2.75-103.php
  [Accessed 27 07 2019].
- Glen, S., 2016. *Statistics How To.* [Online]
  Available at: https://www.statisticshowto.datasciencecentral.com/rmse/
  [Accessed 29 07 2019].
- Glen, S., 2016. *Statistics How To.* [Online]

Available at: https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/correlation-analysis/

[Accessed 29 07 2019].

- goodenergy, n.d. *goodenergy.* [Online]

  Available at: https://www.goodenergy.co.uk/how-do-wind-turbines-work/

- Hansen, M. O. L., 2013. *Aerodynamics of Wind Turbines.* 2 ed. s.l.:Earthscan.

- HaskoningDHV, R., 2017. *Norfolk Vanguard Offshore Wind Farm Non-Technical Summary.* [Online]

  Available at:

  https://corporate.vattenfall.co.uk/globalassets/uk/projects/norfolk-vanguard/consultation-peir-oct-2017/peir/non-technical-summary/non-technical-summary_norfolk-vanguard-peir.pdf

  [Accessed 3 3 2019].

- Hawker, G., 2019. *Yield Assessment and Financing 1.* [Online] [Accessed 10 07 2019].

- Hyndman, R. J., 2016. *Hyndsight blog.* [Online]

  Available at: https://robjhyndman.com/hyndsight/tscv/

  [Accessed 22 07 2019].

- Hyun-Goo Kim, J.-Y. K. Y.-H. K., 2018. Comparative Evaluation of the Third-Generation Reanalysis Data for Wind Resource Assessment of the Southwestern Offshore in South Korea. *Atmosphere,* 9(2).

- I.González-Aparicio & A.Zucker, 2015. Impact of wind power uncertainty forecasting on the market integration of wind energy in Spain. *Applied Energy,* Volume 159, pp. 334-349.

- J. F. Manwell, J. G. M., 2009. *Wind Energy Explained.* 2 ed. Noida¸India: WIley.
- James, G., Witten, D., Hastie, T. & Tibshirani, R., 2013. *An Introduction to Statistical Learning: with Applications in R.* JAMAICA, NY, U.S.A.: Springer.
- Kalkan, A., 2015. *Inores.* [Online]
  Available at:
  https://www.tureb.com.tr/files/iwpc/sunumlar/2nisan/akgun_kalkan.pdf
  [Accessed 09 07 2019].
- Kenton, W., 2019. *Investopedia.* [Online]
  Available at: https://www.investopedia.com/terms/m/mlr.asp
  [Accessed 29 07 2019].
- Lange, M. & Focken, U., 2005. *State-of-the-Art in Wind Power Prediction in Germany and International Developments.* [Online]
  Available at: https://www.researchgate.net/publication/237677945_State-of-the-Art_in_Wind_Power_Prediction_in_Germany_and_International_Developments
  [Accessed 03 06 2019].
- Lobo, M. G., 2010. *Métodos de predicción de la generación agregada de enerrgía eólica.* [Online]
  Available at: https://core.ac.uk/download/pdf/29402021.pdf
- Ltd, B. O. W., 2012. *Beatrice Offshore Windfarm Onshore Transmission Works Environmental Statement.* [Online]

Available at: https://sse.com/media/113175/NonTechnicalSummaryFinal.pdf [Accessed 3 3 2019].

- Manwell, J. F., McGowan, J. G. & Rogers, A. L., 2009. *Wind Energy Explained.* 2 ed. University of Massachusetts, USA: Wiley.
- MarkLandry, T. P. D. C., 2016. Probabilistic gradient boosting machines for GEFCom2014 wind forecasting. *International Journal of Forecasting,* 32(3), pp. 1061-1066.
- McMillan, D., n.d. *Wind Forecasting - Wind Energy and Power System Operation.* [Online].
- Messner, J. W., Mayr, G. J. & Zeileis, A., 2016. Heteroscedastic Censored and Truncated Regression with crch. *R Journal*, 8(1), pp. 173-181.
- Miceli, F., 2017. *windfarmbop.* [Online] Available at: http://www.windfarmbop.com/category/met-mast/ [Accessed 2019].
- Mishra, A., 2018. *Towards Data Science.* [Online] Available at: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234
[Accessed 29 7 2019].
- Models, W. T., 2019. *Siemens SWT-2.3-113.* [Online] Available at: https://en.wind-turbine-models.com/turbines/110-siemens-swt-2.3-113#datasheet
[Accessed 27 7 2019].
- NREL, 1997. *Wind Resource Assessment Handbook: Fundamentals for Conducting a Successful Monitoring.* [Online]

Available at: https://www.nrel.gov/docs/legosti/fy97/22223.pdf [Accessed 19 08 2019].

- Olauson, J. & Bergkvist, M., 2015. *Modelling the Swedish wind power production using MERRA reanalysis data.* [Online]
Available at:
https://www.sciencedirect.com/science/article/pii/S0960148114008167
[Accessed 01 06 2019].
- Patrizio Lecca, Peter G. McGregor, Kim J. Swales, Marie Tamba, 2017. *The Importance of Learning for Achieving the UK's Targets for Offshore Wind.* [Online]
Available at:
https://pureportal.strath.ac.uk/files-asset/64351207/Lecca_etal_EE2017_importance_of_learning_for_achieving_the_UK_s_targets.pdf
- Peck, D. C. M. E. A. & Vining, G. G., 2012. *Introduction to linear regression analysis.* 5 ed. s.l.:Wiley.
- Peter K. Dunn, G. K. S., 2013. *Generalized Linear Models With Examples in R.* Springer Texts in Statistics ed. s.l.:Springer.
- Pierre Pinson, H. A. N. H. M. T. S. N., 2008. *Local Linear Regression with Adaptive Orthogonal Fitting for the Wind Power Application.* [Online]
Available at: DOI: 10.1007/s11222-007-9038-7 · Source: DBLP [Accessed 2019].
- Pitteloud, J.-D., 2019. *Global Wind Installations - Wind Energy International.* [Online]

Available at: https://library.wwindea.org/global-statistics-2018-preliminary/ [Accessed 31 05 2019].

- Press, C., 2008. *Why Wind Turbines Can Mean Death For Bats.* [Online] Available at: https://www.sciencedaily.com/releases/2008/08/080825132107.htm [Accessed 1 3 2019].
- S. M. Lawan, W. A. W. Z. A. W. Y. C. A. B. T. M., 2014. Different Models of Wind Speed Prediction; A Comprehensive Review. *International Journal of Scientific & Engineering Research,* 5(1), pp. 1760-1768.
- Schubel, P. & Crossley, R. J., 2014. *Wind Turbine Blade Design.* [Online] Available at: https://www.researchgate.net/publication/300904446_Wind_Turbine_Blade_Design [Accessed 03 06 2019].
- Shahab Shokrzadeh, M. J. J. E. B., 2014. Wind Turbine Power Curve Modeling Using Advanced Parametric and Nonparametric Methods. *IEEE Transactions on Sustainable Energy,* 5(4), pp. 1262-1269.
- Shaikh, R., 2018. *Towards Data Science.* [Online] Available at: https://towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff85 [Accessed 01 08 2019].
- Singh, G., 2018. *Analytics Vidhya.* [Online] Available at: https://www.analyticsvidhya.com/blog/2018/03/introduction-regression-splines-python-codes/ [Accessed 30 07 2019].

- Smith, B., 2017. *Sciencing.* [Online]

  Available at: https://sciencing.com/advantages-disadvantages-finding-variance-8364027.html

  [Accessed 30 07 2019].
- Soman, S. S., Zareipour, H., Malik, O. & Mandal, P., 2010. *Wind Energy: Forecasting Challenges for Its Operational Management.* [Online] Available at: https://www.researchgate.net/publication/224188805_A_review_of_wind_power_and_wind_speed_forecasting_methods_with_different_time_horizons

  [Accessed 01 06 2019].
- Swalin, A., 2018. *USF-Data Science.* [Online]

  Available at: https://medium.com/usf-msds/choosing-the-right-metric-for-machine-learning-models-part-1-a99d7d7414e4

  [Accessed 29 07 2019].
- *Technical projects in the renewables industry* (2019) Fraser Tough.
- The New Economy, 2018. *The Significant Role Of Wind Energy.* [Online] Available at: http://climatechange-theneweconomy.com/cop23-series-wind-turbines/

  [Accessed 31 05 2019].
- Universidad de Chile, n.d. *Tipos de Generadores y Conversión de la Energía.* [Online]

  Available at: http://hrudnick.sitios.ing.uc.cl/alumno12/viento/17.html
- WeatherSpark, 2019. [Online]

Available at: https://es.weatherspark.com/y/81927/Clima-promedio-en-Hol-Noruega-durante-todo-el-a%C3%B1o
[Accessed 26 07 2019].

- Wind Energy - The Facts, 2019. *Local Wind Resource Assessment and Energy Analysis.* [Online]
  Available at: https://www.wind-energy-the-facts.org/index-11.html
  [Accessed 2019].
- Wind Energy - The Facts, 2019. *The Importance of the Wind Resource.* [Online]
  Available at: https://www.wind-energy-the-facts.org/the-importance-of-the-wind-resource.html
  [Accessed 2019].
- Wind Europe, 2019. *Wind energy in Europe - Trends and statistics.* [Online]
  Available at: https://windeurope.org/wp-content/uploads/files/about-wind/statistics/WindEurope-Annual-Statistics-2018.pdf
  [Accessed 31 05 2019].
- Würth, I. et al., 2019. *Minute-Scale Forecasting of Wind Power—Results from.* [Online]
  Available at: https://www.researchgate.net/publication/331261469_Minute-Scale_Forecasting_of_Wind_Power-Results_from_the_Collaborative_Workshop_of_IEA_Wind_Task_32_and_36
  [Accessed 21 07 2019].
- WWEA, 2019. *Wind Power Capacity Worldwide reaches 600 GW, 53,9 GW added in 2018.* [Online]

Available at: https://wwindea.org/blog/2019/02/25/wind-power-capacity-worldwide-reaches-600-gw-539-gw-added-in-2018/

[Accessed 31 05 2019].

13. Appendices

13.1. Appendix I – Summary of formulas in R language

- for1<- Site~ ERA5+ MERRA+ ConWX + RVM + sinh + cosh
- for2 <- Site~ bs(ERA5,3)+bs(MERRA,3)+bs(ConWX,3)+bs(RVM,3) + sinh + cosh
- for3 <- Site~ ERA5+ MERRA+ ConWX + RVM + sinh4 + cosh4 + sinh + cosh
- for4<- Site~ ERA5+ MERRA + sinh + cosh
- for5 <- Site~ bs(ERA5,3)+bs(MERRA,3) + sinh + cosh
- for6 <- Site~ ERA5+ MERRA + sinh4 + cosh4 + sinh + cosh
- for7<- Site~ ERA5+ ConWX + sinh + cosh
- for8 <- Site~ bs(ERA5,3)+bs(ConWX,3) + sinh + cosh
- for9 <- Site~ ERA5+ ConWX + sinh4 + cosh4 + sinh + cosh
- for10<- Site~ ERA5+ RVM + sinh + cosh
- for11<- Site~ bs(ERA5,3)+bs(RVM,3) + sinh + cosh
- for12 <- Site~ ERA5+ RVM + sinh4 + cosh4 + sinh + cosh
- for13<- Site~ MERRA + RVM + sinh + cosh
- for30 <- Site~ (ERA5- ConWX)
- for31 <- Site~ (ERA5- RVM)
- for32 <- Site~ (MERRA- RVM)
- for33 <- Site~ (ConWX- RVM)
- for34 <- Site~ (MERRA- ConWX)
- for35 <- Site~ bs(ERA5,3)+bs(MERRA,3)+bs(ConWX,3)+bs(RVM,3) + sinh4 + cosh4 + sinh + cosh
- for36 <- Site~ bs(ERA5,3)+bs(MERRA,3)+bs(ConWX,3)+bs(RVM,3) + sinh4 + cosh4 + sinh + cosh + sinh6 + cosh6
- for37 <- Site~ bs(ERA5,3)+ bs(MERRA,3)+ bs(ConWX,3)+ bs(RVM,3)+ sinh + cosh + sinh6 + cosh6
- for38 <- Site~ bs(ERA5,3)+ bs(MERRA,3)+ bs(ConWX,3)+ bs(RVM,3)+ sinh4 + cosh4 + sinh6 + cosh6
- for39 <- Site~ bs(MRT,3) + sinh6 + cosh6
- for40 <- Site~ poly(ERA5,2) +poly(MERRA,2)+ poly(ConWX,2)+

- for14<- Site~ bs(MERRA,3)+bs(RVM,3) + sinh + cosh
- for15 <- Site~ MERRA+ RVM + sinh4 + cosh4 + sinh + cosh
- for16<- Site~ MERRA + ConWX + sinh + cosh
- for17<- Site~ bs(MERRA,3)+bs(ConWX,3) + sinh + cosh
- for18 <- Site~ MERRA+ ConWX + sinh4 + cosh4 + sinh + cosh
- for19<- Site~ RVM + ConWX + sinh + cosh
- for20<- Site~ bs(ConWX,3)+bs(RVM,3) + sinh + cosh
- for21 <- Site~ RVM + ConWX + sinh4 + cosh4 + sinh + cosh
- for22 <- Site~ ERA5+ MERRA + sinh + cosh
- for23 <- Site~ (ERA5- MERRA)+ sinh + cosh
- for24 <- Site~ (ERA5- ConWX)+ sinh + cosh
- for25 <- Site~ (ERA5- RVM)+ sinh + cosh
- for26 <- Site~ (MERRA- RVM)+ sinh + cosh
- for27 <- Site~ (ConWX- RVM)+ sinh + cosh
- for28 <- Site~ (MERRA- ConWX)+ sinh + cosh
- for29 <- Site~ (ERA5- MERRA)

poly(RVM,2)

- for41 <- Site~ poly(ERA5,2)+ poly(MERRA,2)+ poly(ConWX,2)+ poly(RVM,2)+ sinh + cosh
- for42 <- Site~ poly(ERA5,2)+ poly(MERRA,2)+ poly(ConWX,2)+ poly(RVM,2)+ sinh4 + cosh4
- for43 <- Site~ poly(ERA5,2)+ poly(MERRA,2)+ poly(ConWX,2)+ poly(RVM,2)+ sinh4 + cosh4
- for44 <- Site~ poly(ERA5,2)+ poly(MERRA,2)+ poly(ConWX,2)+ poly(RVM,2)+ sinh6 + cosh6
- for45 <- Site~ poly(ERA5,2)+ poly(MERRA,2)+ poly(ConWX,2)+ poly(RVM,2)+ sinh6 + cosh6 + sinh4 + cosh4 +sinh + cosh
- for46 <- Site~ poly(ERA5,2)+ poly(MERRA,2)+ poly(ConWX,2)+ poly(RVM,2)+ sinh6 + cosh6 +sinh + cosh
- for47 <- Site~ poly(ERA5,2)+ poly(MERRA,2)+ poly(ConWX,2)+ poly(RVM,2)+ sinh6 + cosh6 + sinh4 + cosh4
- for48 <- Site~ poly(ERA5,2)+ poly(MERRA,2)+ poly(ConWX,2)+ poly(RVM,2)+ sinh4 + cosh4 +sinh + cosh

13.2. Appendix II - Results

5-fold cross-validation

5fold	*Model 35*					*Model 36*				
Site	BIAS	RMSE	MAE	R2	Correlation	BIAS	RMSE	MAE	R2	Correlation
ENGgra M277	0.002528865	1.051352	0.8152742	0.8534935	0.9139752	0.0025092	1.051956	0.8158247	0.8535838	0.9138638
ENGden M224	0.008583246	1.380387	1.06029	0.8049227	0.8889423	0.008457618	1.380181	1.059701	0.8050625	0.8889685
FRAotr M510	–0.0004973749	1.292731	0.9853689	0.8222081	0.9031407	–0.0005050063	1.292705	0.985451	0.8222442	0.903145
FRApdl M29	–0.001969325	1.74241	1.333317	0.7629835	0.8697077	–0.00193981	1.742251	1.333255	0.7630664	0.869734
GERlac M1	–0.006989778	1.479042	1.152412	0.7968154	0.8366889	–0.007007429	1.14788	1.152308	0.7968572	0.8367491
GERlir M1	–0.006294154	1.418556	1.081809	0.7594137	0.8288142	–0.006421292	1.419196	1.082555	0.7595468	0.8286517

Site	BIAS	RMSE	MAE	R2	Correlation	BIAS	RMSE	MAE	R2	Correlation
GERpr1 M2	0.01567992	1.952474	1.517797	0.76586	0.8642103	0.01579375	1.952804	1.517696	0.7659726	0.8641218
GERpru M1	-0.01842231	1.324957	1.012719	0.83134	0.8654668	-0.01851473	1.325996	1.01324	0.8313649	0.8653208
NIRccg M352	-0.001996661	1.343672	1.039129	0.8476639	0.9205574	-0.0019952	1.343564	1.038964	0.8477132	0.9205713
NIRcgr M78	-0.0007300241	1.513116	1.157348	0.8374607	0.9142368	-0.0007300497	1.513165	1.157348	0.8374671	0.9142313
NIRnid M2	-0.00039091	1.399857	1.071755	0.8706885	0.9307885	-0.0003950697	1.399915	1.071796	0.8706917	0.9307822
NORskv M30	0.01335418	1.900256	1.395766	0.8476555	0.8956209	0.01339302	1.901021	1.396329	0.8476753	0.8955396
NORvar M32	-0.0443854	1.905813	1.413311	0.9106667	0.9250528	-0.04414954	1.906396	1.413957	0.9106703	0.9249766
SCOdun M103	-0.003974983	1.77752	1.339436	0.8862846	0.9271903	-0.003904594	1.777636	1.339454	0.8863167	0.9271704
SCOfre M73	-0.0010187	1.666256	1.288302	0.8231454	0.9072001	-0.001015916	1.66634	1.288411	0.8231477	0.907191
SWErod M39	-0.004522026	1.011159	0.7827795	0.8490414	0.9155016	-0.004511525	1.01103	0.7827163	0.8491198	0.9155196

Site	BIAS	RMSE	MAE	R2	Correlation	BIAS	RMSE	MAE	R2	Correlation
TURcig M41	–0.01045559	2.076666	1.584013	0.7546959	0.8610055	–0.01044185	2.076295	1.583769	0.7547817	0.8610672
TURevr M105	–0.001869131	1.936073	1.46773	0.7282772	0.8552265	–0.001948446	1.934881	1.466862	0.7285522	0.8554161
TURhvz M20	–0.02066067	2.077342	1.582465	0.5859354	0.7671786	–0.02057618	2.077298	1.582303	0.5861433	0.7671818
WALbry M578	–0.0003125738	1.312567	1.014369	0.8402746	0.9143865	–0.0003164957	1.312574	1.0144	0.8402825	0.9143853
WALglh M294	–0.003220049	1.282565	0.9811033	0.8676209	0.9227619	–0.003214867	1.282548	0.981163	0.867633	0.9227647
WALmmr M281	–0.006820559	1.479737	1.134438	0.8231757	0.9060448	–0.006846585	1.480615	1.135123	0.8233541	0.9059188

5fold *Worst model: MODEL 33*

Site	BIAS	RMSE	MAE	R2	Correlation
ENGgra M277	0.001232987	1.295925	0.994828	0.7715745	0.8655232
ENGden M224	0.006147853	1.534876	1.180879	0.7388771	0.8586318
FRAotr M510	-0.0002409723	1.645146	1.269561	0.7113431	0.8375689
FRApdl M29	2.448222e-05	2.010801	1.520676	0.6794128	0.8218388
GERlac M1	-0.02491985	1.66346	1.294071	0.7318386	0.7975489
GERlir M1	-0.01602813	1.578631	1.212231	0.6997754	0.7877665
GERpr1 M2	-0.004728241	2.175181	1.698538	0.7025565	0.8300779
GERpru M1	-0.003886951	1.851004	1.430793	0.6675457	0.7570942

Site	BIAS	RMSE	MAE	R2	Correlation
NIRccg M352	-0.002494987	1.64855	1.281903	0.7714802	0.8777568
NIRcgr M78	-0.0006814328	1.834937	1.412712	0.7629027	0.8712549
NIRnid M2	-0.0008547151	1.715197	1.314637	0.8031562	0.8941011
NORskv M30	0.01191901	2.140518	1.594456	0.7875889	0.8613584
NORvar M32	-0.01476225	2.590437	2.022119	0.8226851	0.8547487
SCOdun M103	-0.01118914	2.085747	1.572519	0.8416472	0.8961721
SCOfre M73	-0.000144947	1.955523	1.504612	0.7601213	0.8706532
SWErod M39	-0.00125983	1.265625	0.979872 2	0.7639148	0.8643913
TURcig M41	-0.001597728	2.382281	1.794984	0.6694211	0.8138813
TURevr M105	-0.003740463	2.132222	1.633396	0.6652523	0.8194285

Site	BIAS	RMSE	MAE	R2	Correlation
TURhvz M20	-0.007180739	2.283836	1.751667	0.4860944	0.7072149
WALbry M578	-0.0003406037	1.555172	1.198194	0.7733775	0.8774438
WALglh M294	-0.006051825	1.570303	1.199825	0.7987897	0.8807149
WALmmr M281	-0.003190619	1.801371	1.396267	0.7332066	0.8576365

6- and 10-fold cross-validation comparison for the best model

Model 36	*6 FOLDS*					*10 FOLDS*				
Site	BIAS	RMSE	MAE	R2	Correlation	BIAS	RMSE	MAE	R2	Correlation
ENGgra M277	0.002857095	1.047828	0.8135081	0.8505836	0.9140445	-8.202271e-05	1.046234	0.8122577	0.8481829	0.9124449
ENGden M224	0.004453406	1.37897	1.058897	0.8020968	0.8851257	0.002899675	1.374811	1.055709	0.8182197	0.8743533
FRAotr M510	-0.000424468	1.29283	0.9853908	0.8201412	0.9031216	-0.0007038368	1.292693	0.9852509	0.8188199	0.9017793
FRApdl M29	-0.00325942	1.741338	1.332577	0.7643214	0.867778	-0.003566412	1.741852	1.333313	0.7624226	0.866
GERlac M1	-0.01290613	1.478855	1.153075	0.796471	0.8392581	-0.006384506	1.469846	1.147486	0.7915845	0.8352286
GERlir M1	-0.005243752	1.401615	1.075221	0.7641089	0.825836	-0.001427101	1.403302	1.074431	0.7580732	0.8195725
GERpr1 M2	0.01495752	1.958031	1.521194	0.7638903	0.8518199	0.00635593	1.937411	1.508801	0.7599653	0.8538337

Site	BIAS	RMSE	MAE	R2	Correlation	BIAS	RMSE	MAE	R2	Correlation
GERpru M1	-0.02081572	1.331374	1.017262	0.830546	0.8660314	-0.01118604	1.320923	1.010557	0.8315843	0.8618804
NIRccg M352	-0.0001359732	1.343304	1.038794	0.8474716	0.9209165	0.0002454331	1.343365	1.038759	0.8480681	0.9155703
NIRcgr M78	-0.00100753	1.511015	1.156098	0.8399567	0.9150402	-0.001352039	1.513213	1.157428	0.8421838	0.907487
NIRnid M2	0.0002570854	1.399369	1.071474	0.869672	0.9311324	-0.000570953	1.399953	1.071832	0.8688497	0.9264006
NORskv M30	0.01598027	1.902129	1.397697	0.8467653	0.8868799	0.0005126885	1.893958	1.390756	0.8361216	0.8795572
NORvar M32	-0.02893649	1.91004	1.415382	0.9058252	0.9224496	-0.008686274	1.886905	1.401989	0.8973563	0.9258672
SCOdun M103	-0.001031563	1.773145	1.338439	0.8854724	0.9305429	-0.003809221	1.775143	1.337685	0.8861367	0.9236509
SCOfre M73	-0.002161217	1.667085	1.288767	0.8241964	0.9069544	-0.001776397	1.665834	1.28811	0.8227552	0.9020142
SWErod M39	-0.00272227	1.012424	0.7838285	0.8487444	0.9108654	-0.004249542	1.010162	0.7819844	0.8477326	0.909258

Site	BIAS	RMSE	MAE	R2	Correlation	BIAS	RMSE	MAE	R2	Correlation
TURcig M41	-0.006729557	2.077457	1.584607	0.7521897	0.8607154	-0.002895401	2.074949	1.583425	0.7474298	0.8564331
TURevr M105	-0.007602974	1.933341	1.465196	0.7416467	0.850197	-0.004583383	1.935354	1.46769	0.7367479	0.8537642
TURhvz M20	-0.01151887	2.070329	1.577288	0.5863669	0.7686648	-0.007141275	2.064244	1.575484	0.6123737	0.759664
WALbry M578	0.00061501	1.314048	1.015512	0.8421207	0.9148431	0.001048274	1.314075	1.015615	0.8399781	0.9131192
WALglh M294	0.001511263	1.28254	0.9813373	0.8673936	0.9219812	-0.0003458732	1.28094	0.9801426	0.8647706	0.920935
WALmmr M281	-0.009992744	1.479844	1.134406	0.8255147	0.905074	-0.006688157	1.479119	1.133986	0.8288341	0.902716

Models used with the mean of all the reanalysis tools

	Best Model: Model 5					*Worst Model: Model 12*				
Site	BIAS	RMSE	MAE	R2	Correlation	BIAS	RMSE	MAE	R2	Correlation
ENGgra M277	0.002775831	0.9006829	0.6948119	0.8906971	0.9381785	0.002334896	0.9008832	0.6954776	0.8897169	0.9379368
ENGden M224	0.008901025	1.120459	0.8638929	0.8684873	0.9285452	0.006480203	1.123662	0.8649785	0.8650875	0.927767
FRAotr M510	–0.0005656798	1.065363	0.811173	0.8787957	0.9353188	–0.00031228	1.081112	0.8250007	0.8748029	0.9333481
FRApdl M29	–0.001122165	1.391988	1.067633	0.8467071	0.9187781	–0.0002443245	1.399928	1.074068	0.8441666	0.9178782
GERlac M1	–0.007957384	1.120562	0.8658906	0.8779985	0.9084924	–0.007603156	1.13285	0.872006	0.8747455	0.9073914
GERlir M1	–0.004794631	1.103279	0.8451062	0.8530363	0.8967866	–0.007635981	1.103278	0.8488556	0.8518375	0.8972739

Site	BIAS	RMSE	MAE	R2	Correlation	BIAS	RMSE	MAE	R2	Correlation
GERpr1 M2	0.006944882	1.391725	1.083262	0.8745424	0.9327281	0.001868105	1.405011	1.095446	0.8713945	0.9309371
GERpru M1	0.005003987	1.11567	0.8624382	0.8776055	0.9099346	0.005850234	1.134064	0.8760506	0.8711499	0.9047531
NIRccg M352	-0.0006773408	1.109703	0.8567796	0.8963487	0.9466977	-0.0003468827	1.109815	0.8572803	0.8962545	0.9467137
NIRcgr M78	-0.0003314909	1.232931	0.9503267	0.8916278	0.9437158	-0.0008041765	1.239571	0.9546901	0.8906775	0.9432191
NIRnid M2	0.0002254314	1.171867	0.8943095	0.9091561	0.9521604	-0.0001078286	1.172614	0.8950772	0.908719	0.9520853
NORskv M30	0.006386823	1.393754	1.028855	0.9150242	0.9438572	0.005546596	1.394532	1.030154	0.9133997	0.9434786
NORvar M32	0.001395472	1.515502	1.116123	0.9422718	0.955594	0.008231753	1.509762	1.113459	0.941741	0.9555828
SCOdun M103	-0.00116526	1.326813	1.00423	0.935014	0.9597673	-0.004032767	1.327509	1.003419	0.9347507	0.9597947
SCOfre M73	-9.915207e-05	1.355843	1.048497	0.8829498	0.9395837	7.541166e-05	1.356664	1.049299	0.8825808	0.9395046

Site	BIAS	RMSE	MAE	R2	Correlation	BIAS	RMSE	MAE	R2	Correlation
SWErod M39	0.001136067	0.8823529	0.6813962	0.8853633	0.936345	-0.000570253	0.8892498	0.6856494	0.8833287	0.9353439
TURcig M41	-0.006266717	1.507053	1.148519	0.8687118	0.9290775	-0.001208253	1.521337	1.160112	0.865247	0.9274709
TURevr M105	-0.0004910995	1.466099	1.109518	0.8415375	0.9197094	-0.001830225	1.489112	1.133346	0.8356259	0.9168871
TURhvz M20	0.01114384	1.514446	1.153356	0.7773	0.8831946	-0.00246526	1.49732	1.14721	0.775587	0.8848446
WALbry M578	0.0006746581	1.094887	0.8414453	0.8891024	0.9416628	-0.0002013035	1.100055	0.8457772	0.8875852	0.9410984
WALglh M294	-0.002742816	1.052173	0.8068253	0.9101899	0.9484078	-0.003789549	1.060906	0.8145431	0.9075871	0.9477893
WALmmr M281	-0.001591154	1.151294	0.8891493	0.891146	0.9445503	-0.0007129613	1.153591	0.8918963	0.8900533	0.9443622

Backward cross-validation iteration 1

Model 36 *Iteration 1*

Site	BIAS	RMSE	MAE	R2	Correlation	Training length	Testing length
ENGgra M277	0.06182642	1.048725	0.8144492	0.8662356	0.91545837	4429	6551
ENGden M224	0.18521143	1.355536	1.0397687	0.7858919	0.91004712	3024	897
FRAotr M510	-0.08062451	1.275716	0.9724612	0.8283262	0.87799777	27665	7028
FRApdl M29	-0.05484661	1.771017	1.3715593	0.7549403	0.87758789	5160	5038
GERlac M1	0.08136051	2.414170	1.5684318	0.6177428	0.73795626	1817	4515
GERlir M1	-2.48316964	8.096747	3.8000896	0.9062263	-0.34299612	63	308

BIAS	RMSE	MAE	R2	Correlation	Training length	Testing length	BIAS
GERpr1 M2	0.23247005	1.870869	1.4922833	0.7449620	0.77686866	7561	376
GERpru M1	-0.20337771	1.734639	1.2976393	0.7140781	0.53497205	377	283
NIRccg M352	-0.11883236	1.360706	1.0450249	0.8547279	0.90007143	26474	2314
NIRcgr M78	-0.05879001	1.484023	1.1358500	0.8393586	0.90394638	16086	4730
NIRnid M2	-0.02239385	1.425334	1.0811376	0.8472362	0.92860531	5246	20220
NORskv M30	0.28463380	1.428409	1.1127922	0.8110261	0.87185618	2737	2579
NORvar M32	-0.01376786	1.314981	1.0107322	0.8982791	0.94015599	2078	774
SCOdun M103	0.10328312	1.855462	1.3900613	0.9004612	0.93153220	4473	5020
SCOfre M73	0.04940184	1.678738	1.3034302	0.8379415	0.90055614	16424	15528

BIAS	RMSE	MAE	R2	Correlation	Training length	Testing length	BIAS
SWErod M39	1.04188643	5.231028	2.6564780	0.6981269	-0.05673556	106	7682
TURcig M41	0.07647497	2.110091	1.5822682	0.8133742	0.83443564	5313	4849
TURevr M105	-0.07485176	1.963595	1.4833863	0.7149356	0.82325592	6362	5698
TURhvz M20	-0.49413476	2.071736	1.5564767	0.6180733	0.72064442	1683	4990
WALbry M578	-0.09936856	1.347502	1.0332070	0.8520165	0.90847392	15697	6732
WALglh M294	-0.47117238	1.475378	1.1236295	0.7876610	0.92087551	2370	19436
WALmmr M281	-0.26545882	1.465068	1.1117919	0.8369204	0.91517385	2296	3382

Backward cross-validation iteration 2

Model 36 *Iteration 2*

Site	BIAS	RMSE	MAE	R2	Correlation	Training length	Testing length
ENGgra M277	0.02700271	1.144074	0.8557655	0.8580825	0.9040867	1114	5045
ENGden M224	0.07950555	1.337504	1.026978	0.8332033	0.8914762	4772	2483
FRAotr M510	-0.0650047	1.295561	0.9826312	0.8240383	0.8966841	38251	6071
FRApdl M29	– 0.07939894	1.709128	1.311773	0.7593096	0.8705348	12525	15809
GERlac M1	-0.1938498	1.465224	1.152948	0.7395885	0.7138831	3226	1298
GERlir M1	– 0.00304398	1.345476	1.031558	0.744646	0.8842528	3487	3255
GERpr1 M2	0.681578	2.860182	2.099432	0.8615686	0.7814341	302	5753

Site	BIAS	RMSE	MAE	R2	Correlation	Training length	Testing length
GERpru M1	-0.5214247	1.452462	1.067587	0.8020812	0.9142957	1788	822
NIRccg M352	0.0439089	1.343559	1.046783	0.8329967	0.9286844	12302	12900
NIRcgr M78	-0.1192146	1.390228	1.059705	0.8390514	0.9020519	14939	1649
NIRnid M2	-0.109443	1.431162	1.086894	0.8712275	0.9319812	17734	14359
NORskv M30	-0.09904026	1.862193	1.368756	0.8668946	0.906178	1633	4758
NORvar M32	-0.01018693	5.971422	4.866319	0.9185023	0.2830074	45	453
SCOdun M103	-0.1478576	1.599233	1.223289	0.8765169	0.9338542	4725	623
SCOfre M73	0.0301963	1.46961	1.151755	0.8166488	0.8942263	18228	2192
SWErod M39	0.06053092	0.9400583	0.7300726	0.8619899	0.8811379	5807	841

Site	BIAS	RMSE	MAE	R2	Correlation	Training length	Testing length
TURcig M41	-0.1163565	2.250711	1.690572	0.8021131	0.7924664	6636	1109
TURevr M105	-0.07704549	1.987549	1.502095	0.7506023	0.8434608	14576	12066
TURhvz M20	-0.1179281	1.769403	1.344583	0.6422155	0.742885	3478	1186
WALbry M578	0.1922963	1.420767	1.087134	0.8364519	0.911753	2181	6797
WALglh M294	0.01399584	1.101505	0.8584282	0.885014	0.910624	13256	5962
WALmmr M281	0.01487057	1.635589	1.265431	0.8485595	0.8861552	1276	695

Backward cross-validation iteration 3

Model 36 *Iteration 3*

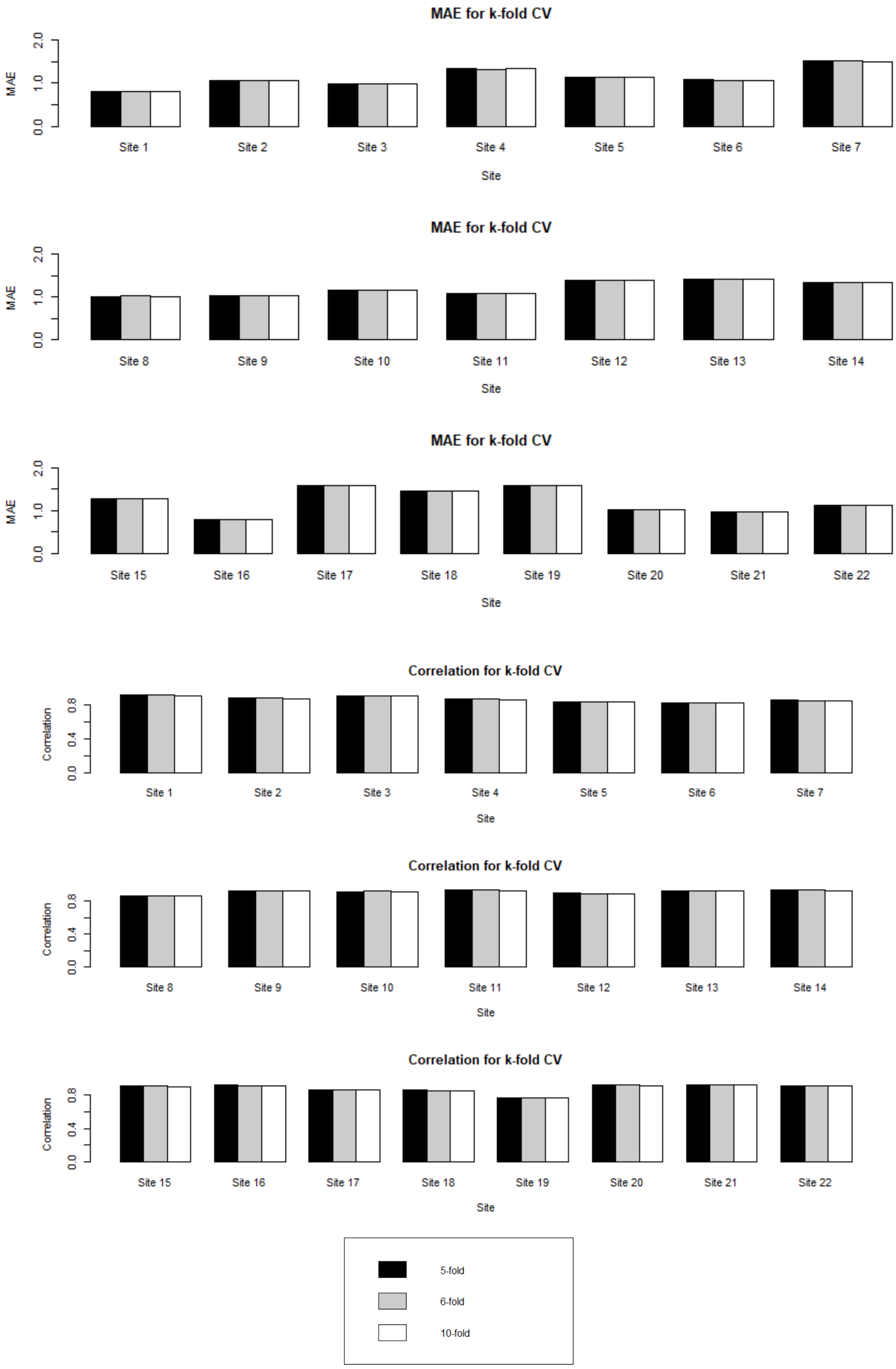
Site	BIAS	RMSE	MAE	R2	Correlation	Training length	Testing length
ENGgra M277	0.25841452	1.133083	0.8786231	0.8562031	0.9212563	1366	4341
ENGden M224	0.10221593	1.352681	1.0507031	0.8112891	0.8888864	956	2166
FRAotr M510	-0.04096578	1.268623	0.9779087	0.8136266	0.9302599	23831	1752
FRApdl M29	0.01639916	1.774158	1.3928273	0.7558991	0.7035405	13369	73
GERlac M1	-0.33571238	2.594011	1.5477783	0.5793273	0.1789955	194	3013
GERlir M1	-26.78497458	71.710102	37.6413022	0.8054692	-0.3807553	31	3799

Site	BIAS	RMSE	MAE	R2	Correlation	Training length	Testing length
GERpr1 M2	–0.08880560	2.067301	1.6285930	0.7245372	0.8725241	1480	2172
GERpru M1	–0.76974499	1.954621	1.3717365	0.7159777	0.8370237	3239	2808
NIRccg M352	–0.09985635	1.343540	1.0376839	0.8619290	0.9218756	10163	25198
NIRcgr M78	0.04706435	1.437672	1.0973066	0.8381406	0.9256202	13784	5511
NIRnid M2	–0.19346915	1.425683	1.0665575	0.8697569	0.8566567	5558	1489
NORskv M30	–0.26479373	2.027256	1.4679255	0.8640144	0.8977605	1437	4050
NORvar M32	0.26016919	2.928205	2.1875169	0.9140720	0.9162450	906	655
SCOdun M103	0.16311968	1.865103	1.3855161	0.8564223	0.9361338	1913	8715
SCOfre M73	0.48090803	1.718108	1.3515027	0.8377683	0.9165008	11268	2708

Site	BIAS	RMSE	MAE	R2	Correlation	Training length	Testing length
SWErod M39	-0.10736118	1.071544	0.8376026	0.8490871	0.9234342	13699	2933
TURcig M41	-0.39720245	2.230375	1.7270936	0.8455338	0.8049952	1518	7256
TURevr M105	0.62979384	1.734604	1.3671737	0.7136580	0.8884581	1889	129
TURhvz M20	0.13200816	1.934800	1.5014051	0.6156447	0.7763803	11756	3879
WALbry M578	0.04901661	1.271998	0.9850143	0.8530165	0.9100164	18666	13376
WALglh M294	-0.09800852	1.217821	0.9199974	0.8634734	0.9444574	13330	3598
WALmmr M281	-0.04260814	1.589535	1.2203326	0.8845419	0.9136900	1157	5163

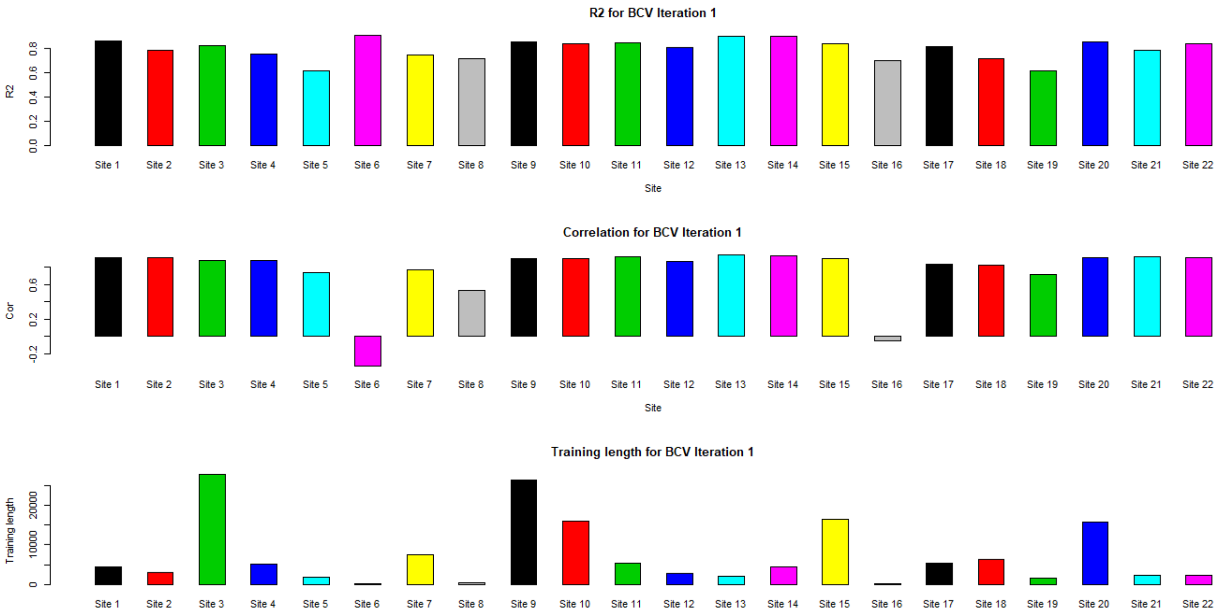
13.3. Appendix III – Graphs

Comparison of MAE and correlation in 5-, 6- and 10- fold CV

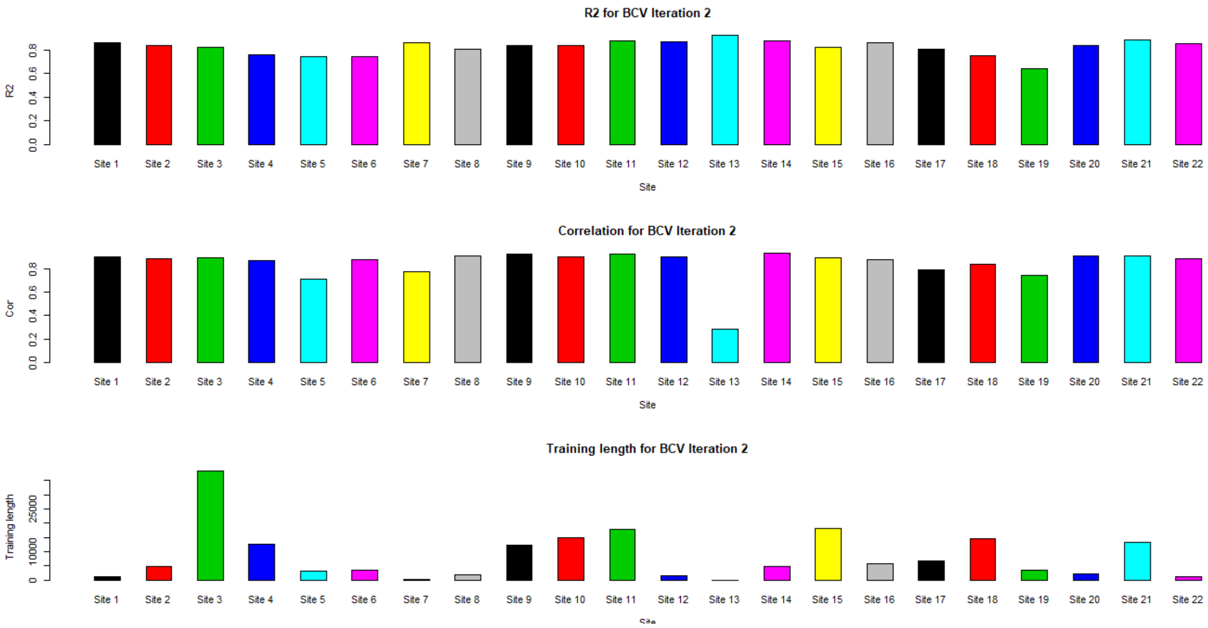


Comparison of R2, correlation and training length in BCV

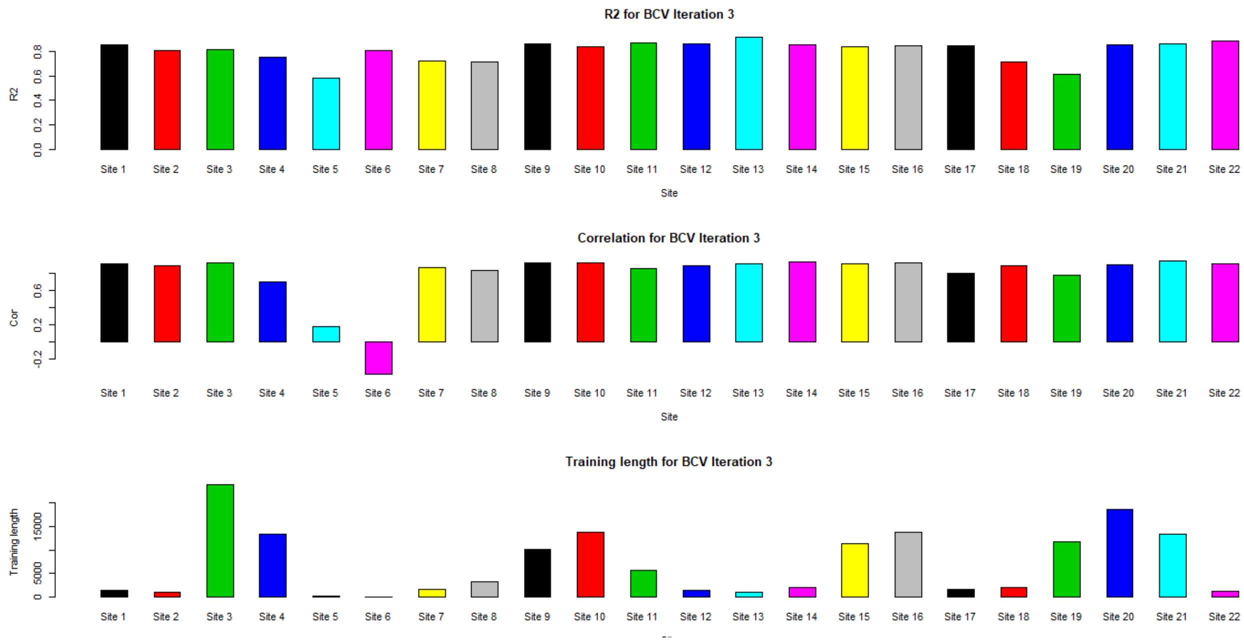
Iteration 1



Iteration 2

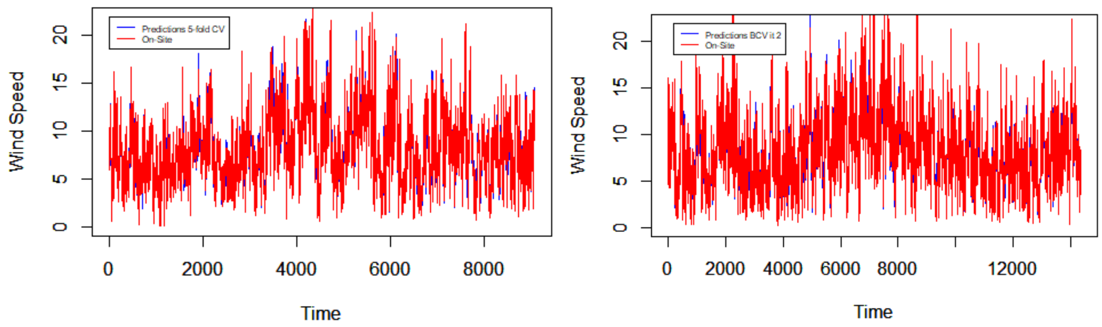


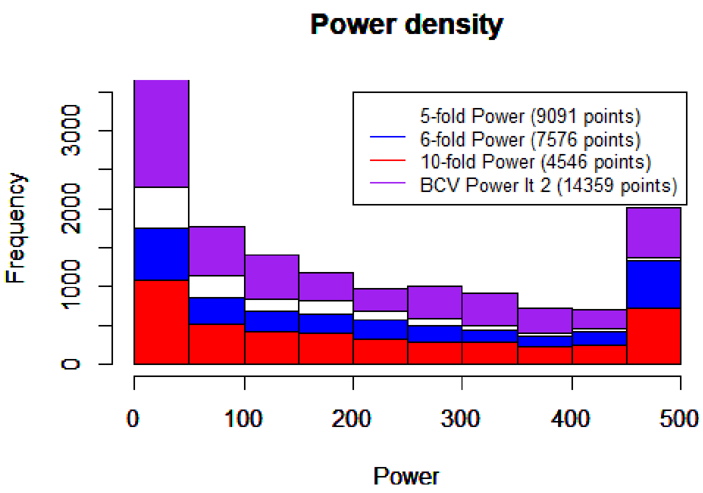
Iteration 3



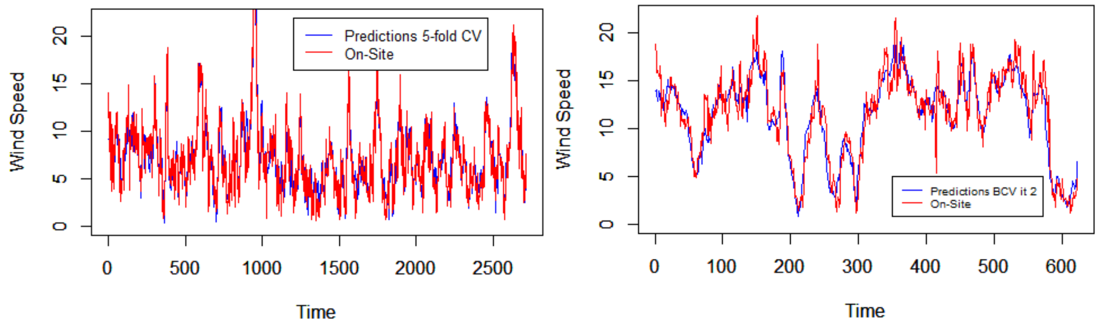
Comparison between predictions and power density (Rated power order)

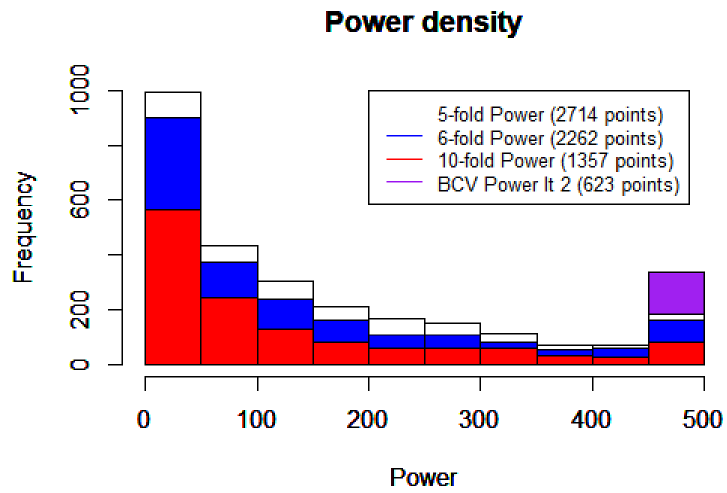
Site 11



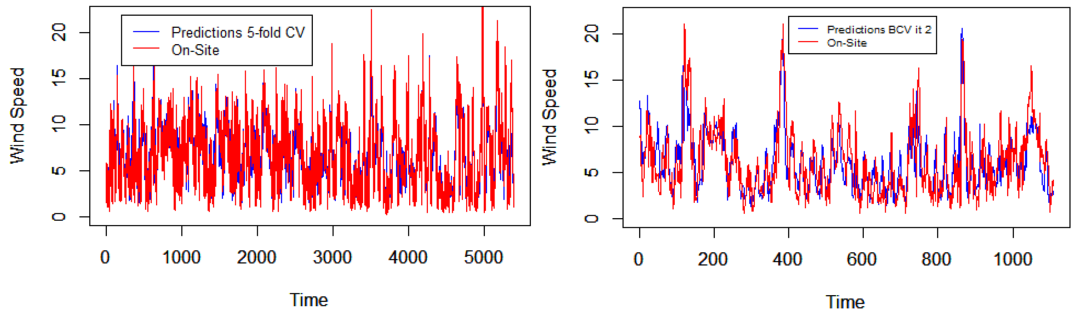


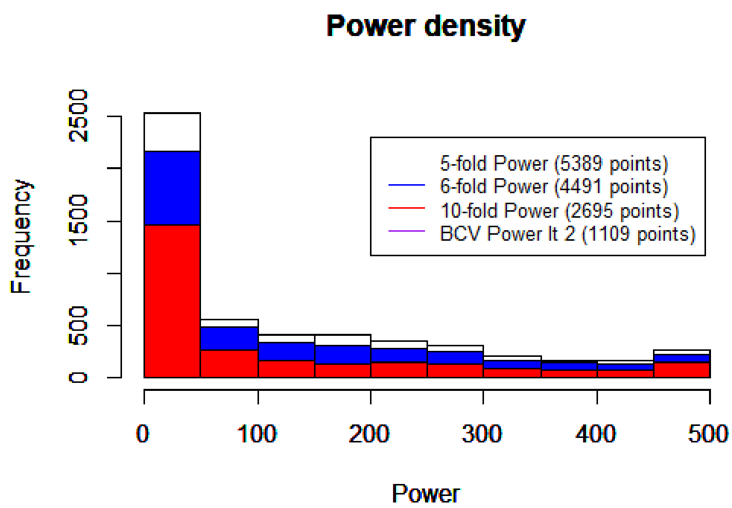
Site 14



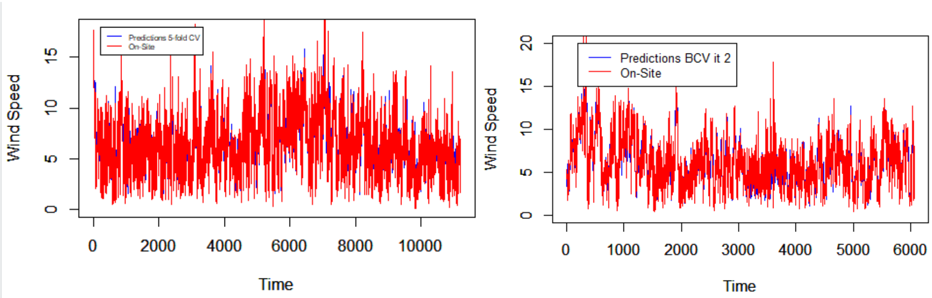


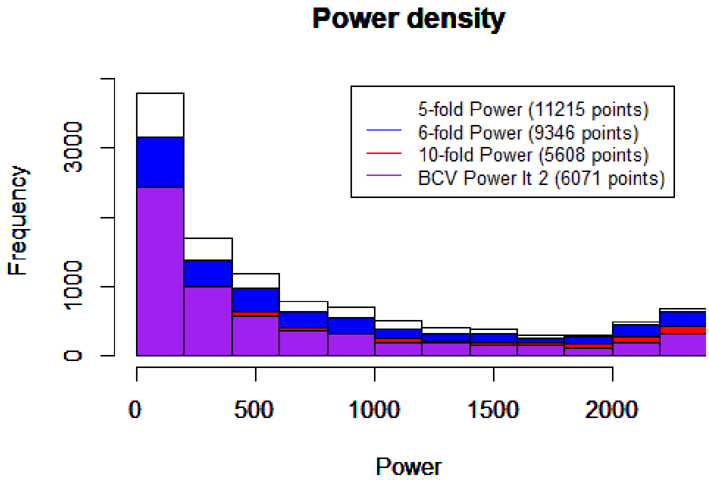
Site 17



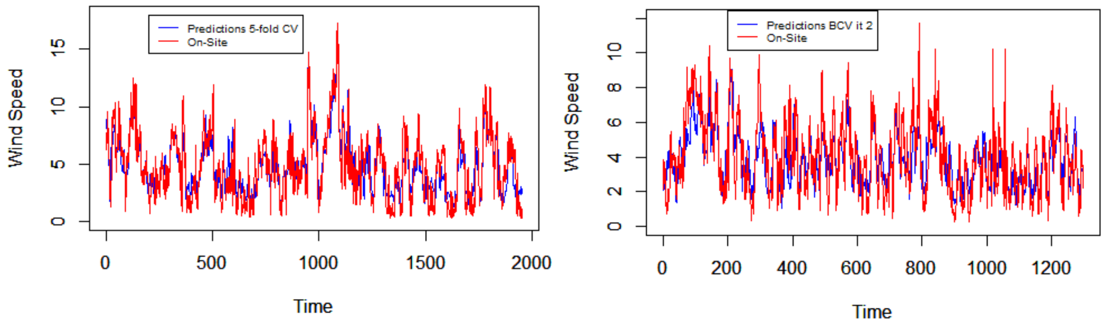


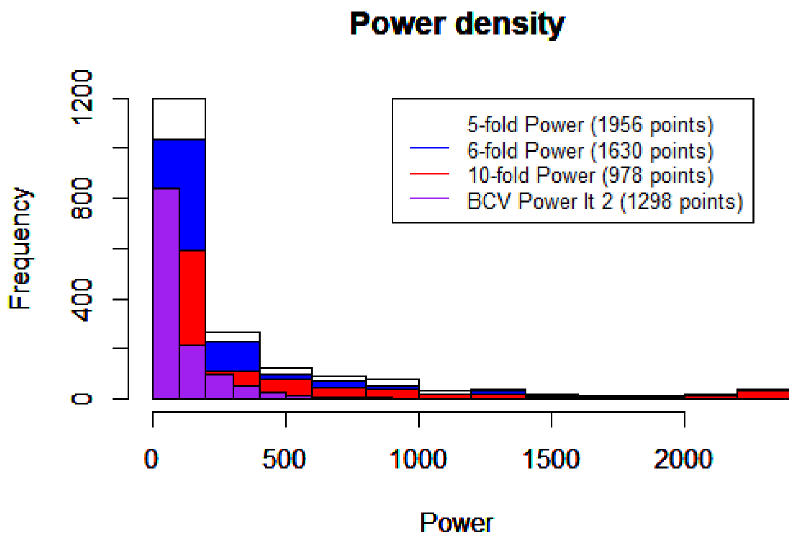
Site 3


Predictions 5-fold CV
On-Site
Wind Speed
Time
Predictions BCV it 2
On-Site
Wind Speed
Time

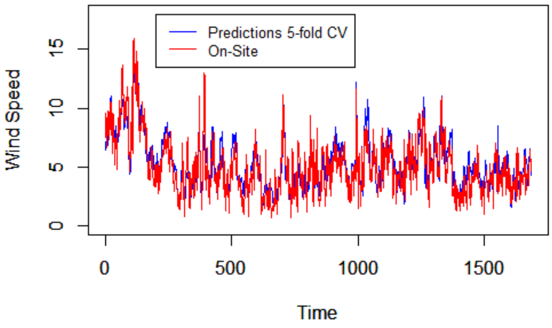

Power density
5-fold Power (11215 points)
6-fold Power (9346 points)
10-fold Power (5608 points)
BCV Power It 2 (6071 points)
Frequency
Power

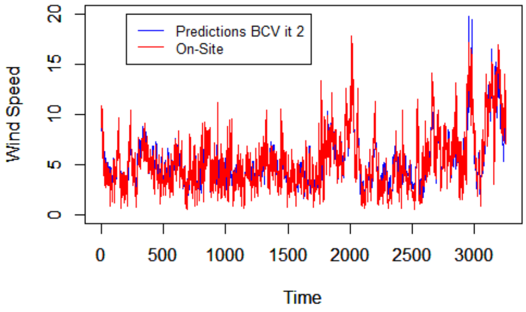
Site 5

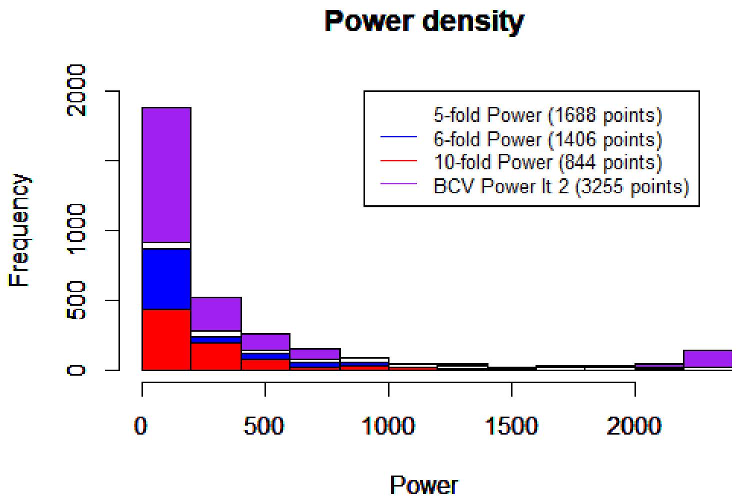

Predictions 5-fold CV
On-Site
Wind Speed
Time
Predictions BCV it 2
On-Site
Wind Speed
Time


Power density
5-fold Power (1956 points)
6-fold Power (1630 points)
10-fold Power (978 points)
BCV Power It 2 (1298 points)
Frequency
Power

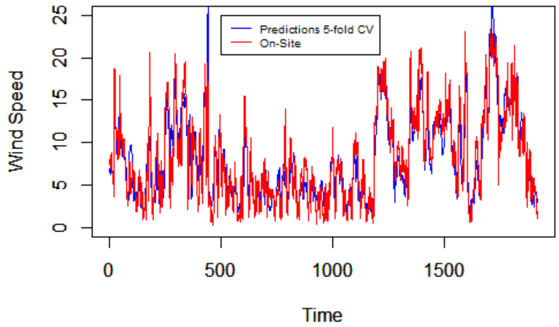
Site 6

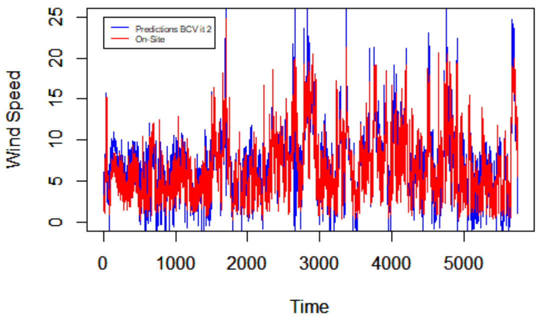

Predictions 5-fold CV
On-Site
Wind Speed
Time

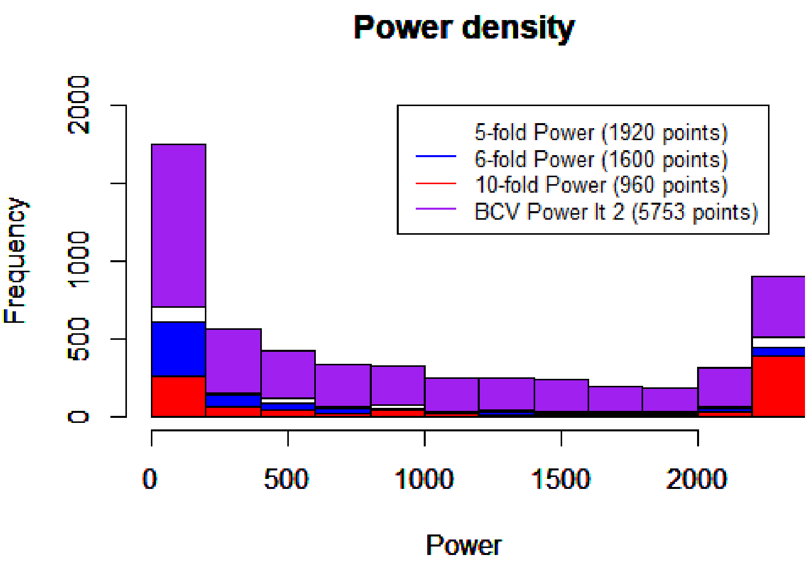

Predictions BCV it 2
On-Site
Wind Speed
Time


Power density
5-fold Power (1688 points)
6-fold Power (1406 points)
10-fold Power (844 points)
BCV Power It 2 (3255 points)
Frequency
Power

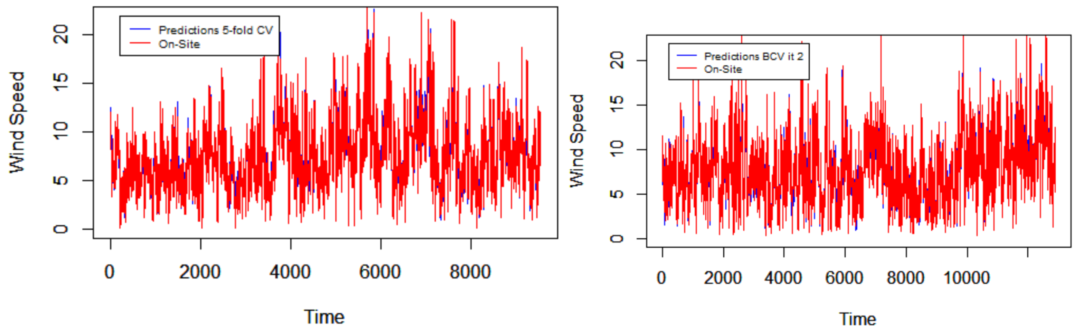
Site 7

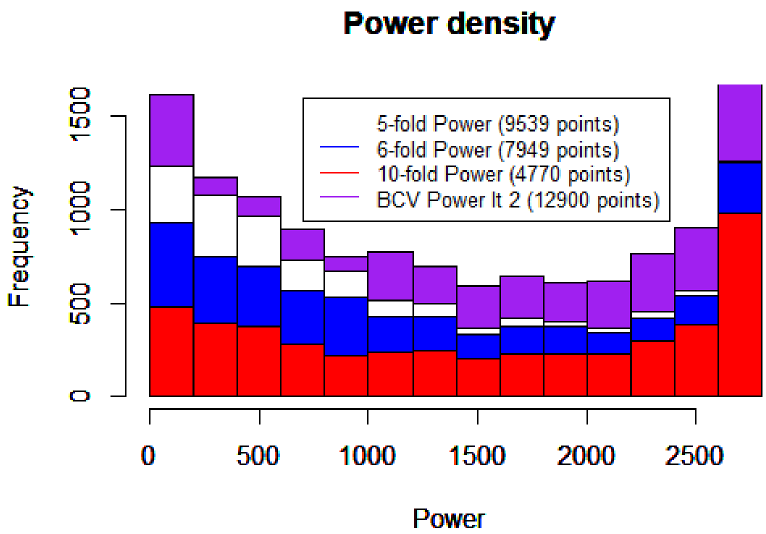

Predictions 5-fold CV
On-Site
Wind Speed
Time


Wind Speed
Time

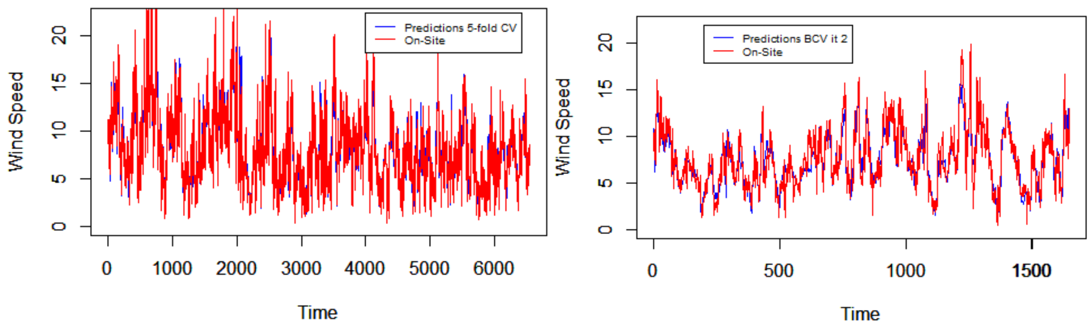

Power density
5-fold Power (1920 points)
6-fold Power (1600 points)
10-fold Power (960 points)
BCV Power It 2 (5753 points)
Frequency
Power

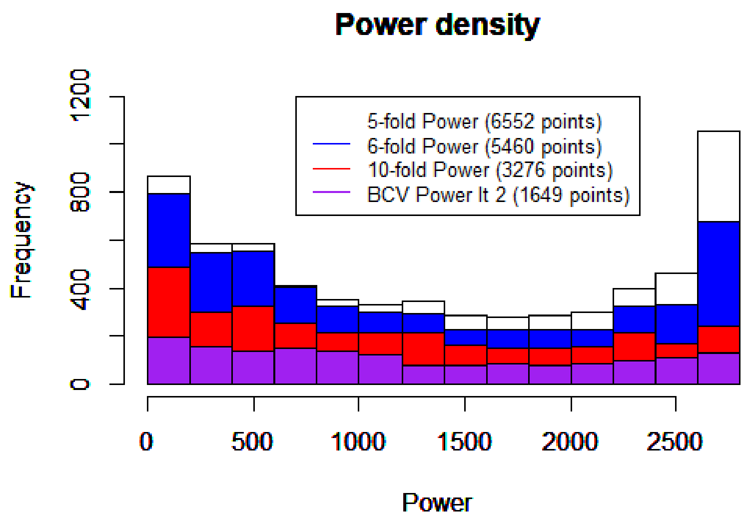
Site 9



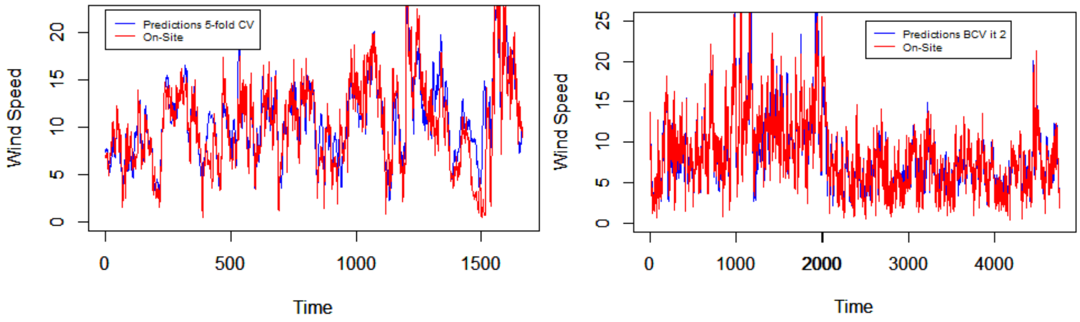


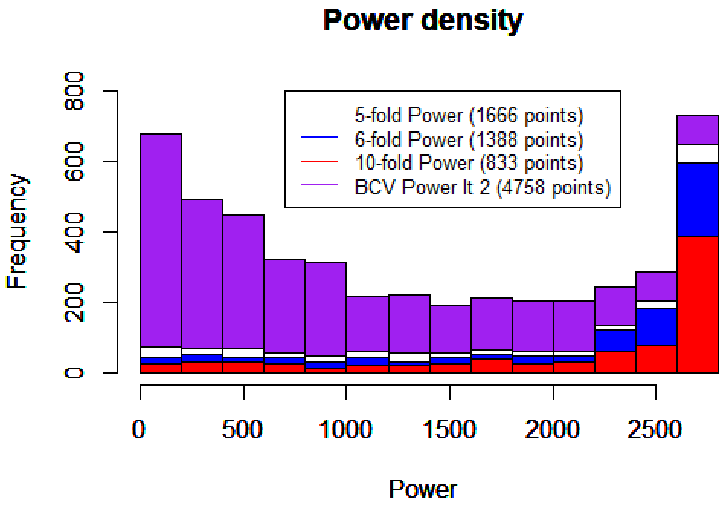
Site 10



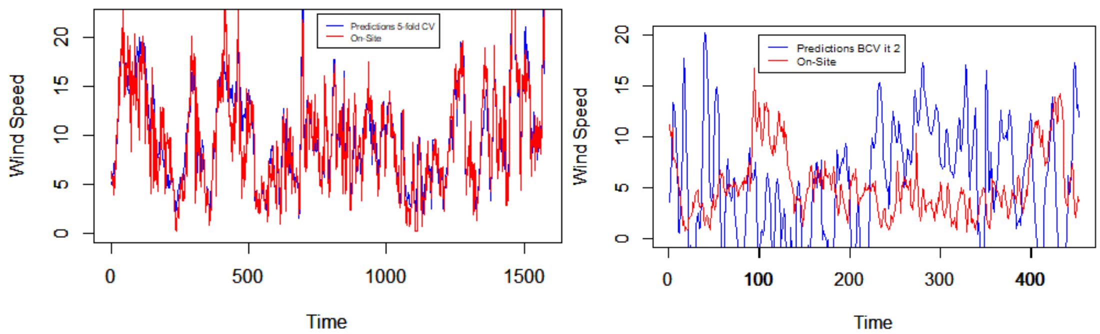


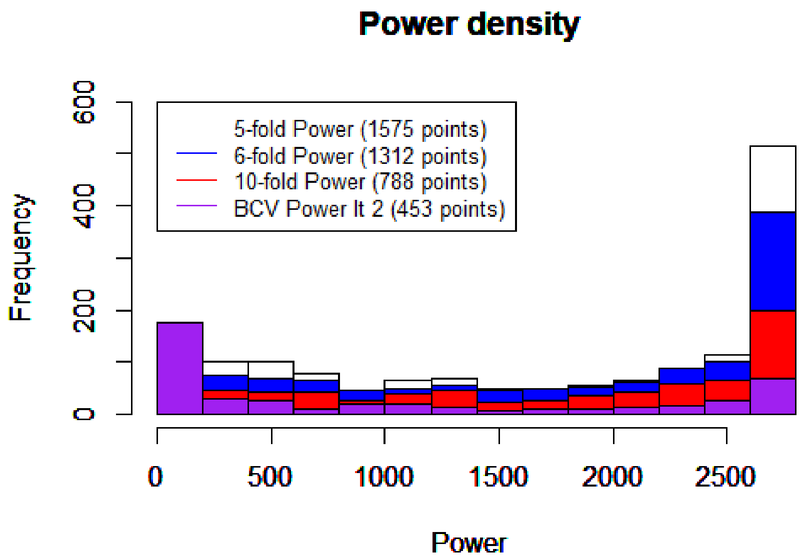
Site 12



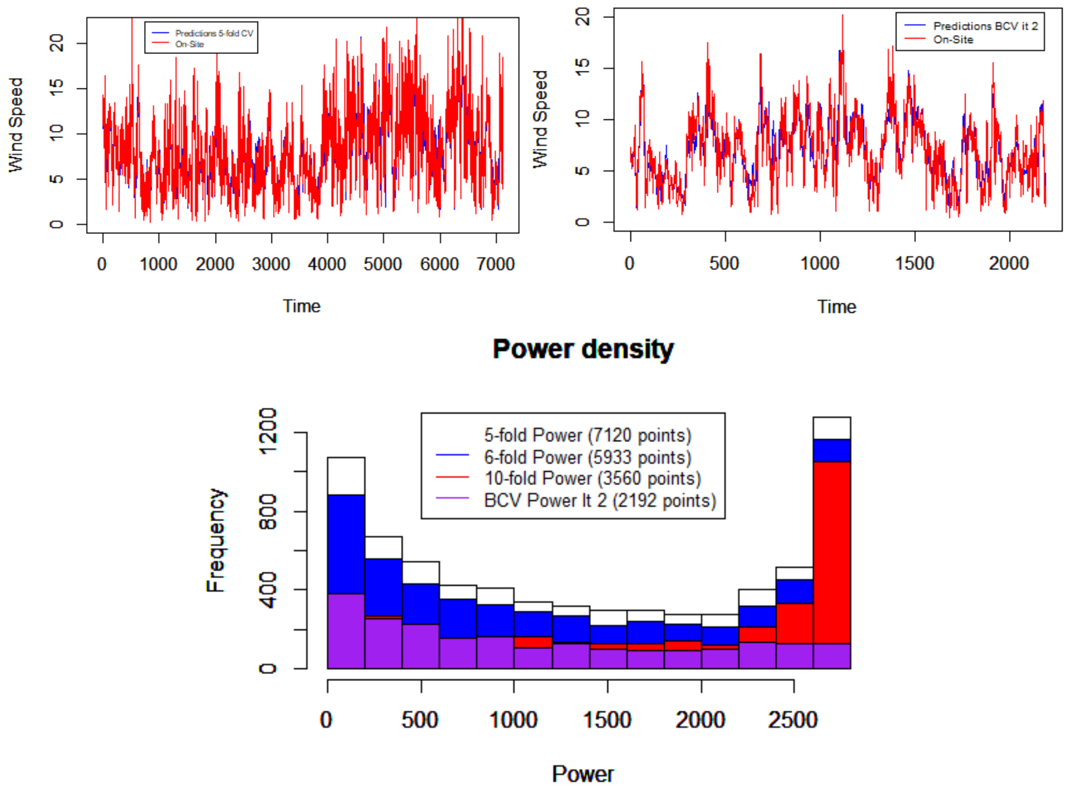


Site 13

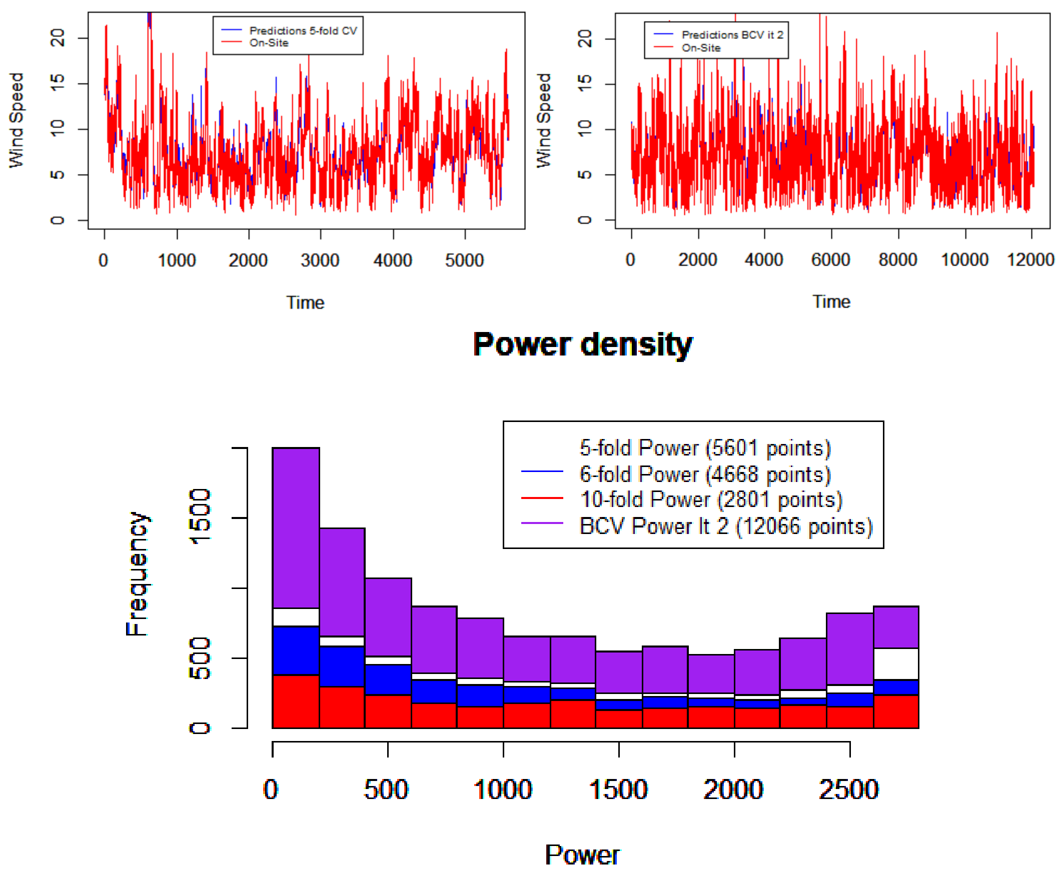




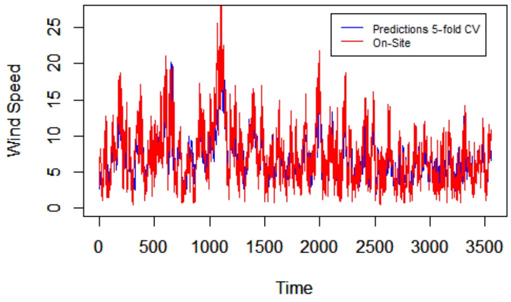
Site 15

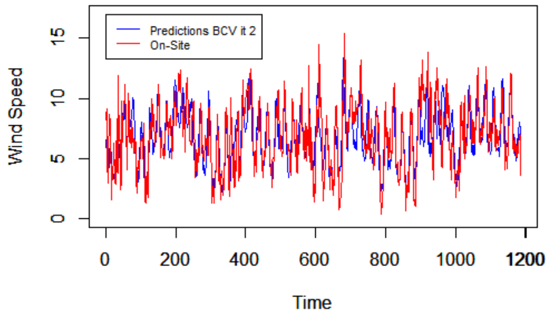


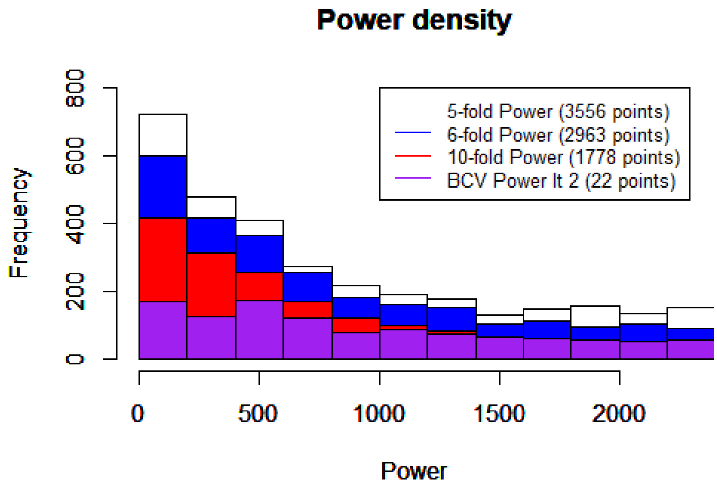
Site 18



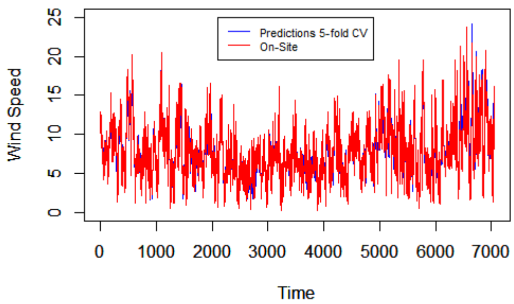
Site 19

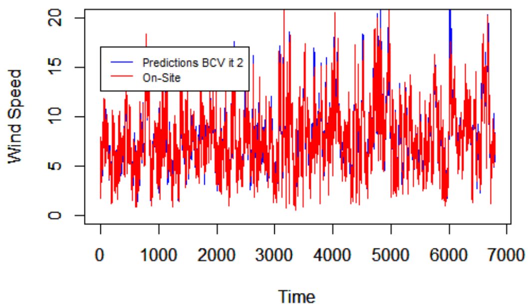


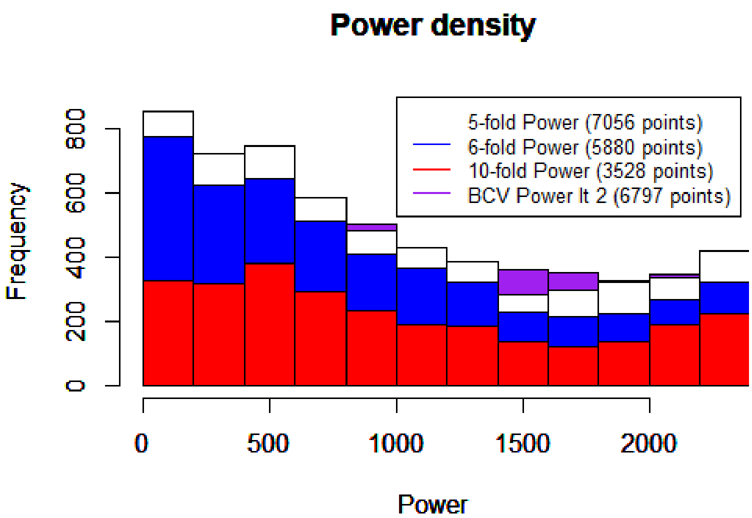




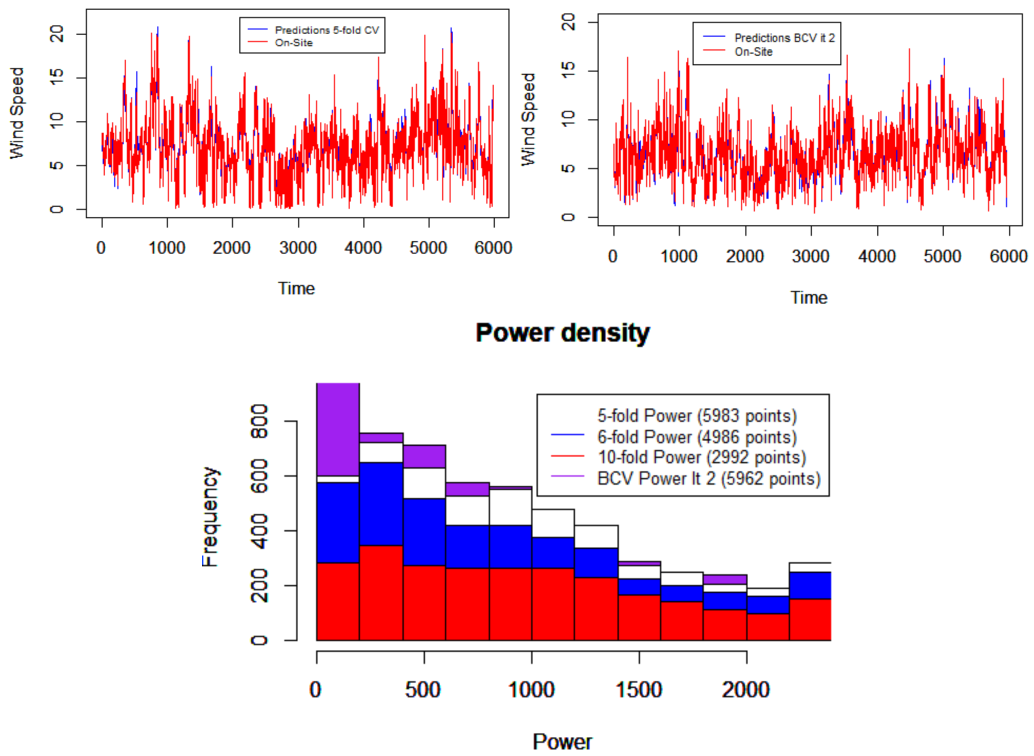
Site 20







Site 21



Site 22