Forecasting Maximum Wind Speed at Offshore Sites

Baptiste Letellier
Abstract

For energy companies involved in the construction and operation of offshore wind power plants, such as Vattenfall, the knowledge of maximum wind speed is critical for logistics, safety and economic reasons. This thesis investigates the possibility to forecast maximum wind speeds at offshore sites, studies the accuracy of these forecasts, and details the employed methodology, so that it can be adapted to other cases in the future.

In order to produce maximum wind speed forecasts, different statistical models were selected, some of them appropriate for short-term predictions (from 1h to 6h ahead), the others aiming at longer-term predictions (days ahead, up to 72h). The methodology consisted in selecting the right parameters for each model, depending on wind measurements and weather forecasts at the tested sites. Then forecasts were issued using the models' equations, forecasted maximum wind speeds were compared with the real values, and the model delivering the best forecasts selected.

The study demonstrated that, if appropriate statistical models were chosen – such as Vector Auto-Regression for short-term, and Generalized Additive Model for long-term – the average errors of precision for maximum wind speeds prediction were lower than 2 m/s, making the forecasts accurate enough to be used. Some work on the models still has to be done before they can be fully integrated into Vattenfall’s in-house weather forecasting system, but the first results are promising.
SAMMANFATTNING

Energiföretag som Vattenfall, som deltar i byggandet och driften av vindkraftverk, behöver löpande ha korrekt kunskap om vindhastighetsvärden, som är avgörande för logistik, säkerheten och ekonomin. Detta examensarbete undersöker möjligheten att förutsäga maximala vindhastigheter på offshore platser, studerar riktigheten i dessa prognoser, och specificerar tillgängliga och beprövade metoder för väderprognoser så att de kan anpassas till Vattenfalls behov.

För att komma fram med prognoser om maximalt vindhastighet, olika statistiska modeller har valts. Några av dem är lämpliga för kortsiktiga prognoser (från 1 timme och upp till 6 timmar i förväg), andra syftar till att leverera förutsägelser på längre sikt (dagar framåt, upp till 72h). Metodiken bestod i att välja rätt parametrar för varje modell, beroende på vindmätningar och väderprognoser på de testade platserna. Prognoserna utfärdades med hjälp av modellernas ekvationer och de prognostiserade maximala vindhastigheterna jämfördes med de uppmätta värden i verkligheten, därmed kunde den mest lämpade modellen identifieras.

Studien visar att i förekommande fall då statistiska modeller valdes - såsom Vector Auto-Regression för kortsiktiga, och respektive Generaliserad Additiv Modell för långsiktiga prognoser - det genomsnitliga precisionsfelet för förutsägelsen av maximal vindhastighet var lägre än 2 m/s, således var prognoserna tillräckligt noggranna för att appliceras i riktiga tillämpningar. En del arbete på modellernas inlärning återstår att göras innan de kan integreras fullt ut i Vattenfalls interna väderprognossystemet, men de första resultaten härmed är mycket lovande.
# Contents

List of Figures .................................................................................................................................................. iii
List of Tables ...................................................................................................................................................... iv
Abbreviations ..................................................................................................................................................... v
ACKNOWLEDGMENTS ......................................................................................................................................... vi

1 INTRODUCTION ........................................................................................................................................... 1
  1.1 Background ........................................................................................................................................... 1
  1.2 Objectives ............................................................................................................................................. 3
  1.3 Methodology ......................................................................................................................................... 3

2 WIND FORECASTING STATE OF THE ART ............................................................................................. 4
  2.1 Wind forecasting timescales ............................................................................................................. 4
  2.2 Wind forecasting approaches .......................................................................................................... 4
  2.3 Wind forecasting applications ......................................................................................................... 7

3 STATISTICAL METHODS .......................................................................................................................... 8
  3.1 Statistical Modeling Principles ......................................................................................................... 8
  3.2 Reference Models ............................................................................................................................... 9
    3.2.1 Climatology ............................................................................................................................ 9
    3.2.2 Persistence ............................................................................................................................. 9
  3.3 Short-Term Forecasting Models ......................................................................................................... 10
    3.3.1 Drift method ............................................................................................................................ 10
    3.3.2 Exponential smoothing .......................................................................................................... 10
    3.3.3 ARIMA models ....................................................................................................................... 10
    3.3.4 Linear and quantile regression .............................................................................................. 12
    3.3.5 VAR models ........................................................................................................................... 13
  3.4 Long-Term Forecasting Models .......................................................................................................... 15
    3.4.1 Linear Regression .................................................................................................................... 15
    3.4.2 Generalized Additive Models (GAM) .................................................................................. 15

4 EXPERIMENTAL DATA ............................................................................................................................. 17
  4.1 Presentation of data ............................................................................................................................. 17
  4.2 Analysis of data .................................................................................................................................. 20

5 METHODOLOGY ........................................................................................................................................... 23
  5.1 General process ................................................................................................................................. 23
  5.2 MAE, RMSE and Skill Score ............................................................................................................. 24

6 RESULTS ....................................................................................................................................................... 25
  6.1 Short-Term Forecasting ..................................................................................................................... 25
    6.1.1 Initial Assessment ...................................................................................................................... 25
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1.2</td>
<td>Results and Analysis</td>
<td>27</td>
</tr>
<tr>
<td>6.1.3</td>
<td>Forecasts versus reality</td>
<td>29</td>
</tr>
<tr>
<td>6.2</td>
<td>Long-Term Forecasting</td>
<td>32</td>
</tr>
<tr>
<td>6.2.1</td>
<td>Initial Assessment</td>
<td>32</td>
</tr>
<tr>
<td>6.2.2</td>
<td>Results and Analysis</td>
<td>33</td>
</tr>
<tr>
<td>6.2.3</td>
<td>Forecasts versus reality</td>
<td>41</td>
</tr>
<tr>
<td>6.3</td>
<td>Maximum Forecasting Error</td>
<td>44</td>
</tr>
<tr>
<td>7</td>
<td>DISCUSSION AND CONCLUSION</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>REFERENCES</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>APPENDIX 1: ADDITIONAL RESULTS PLOTS</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>APPENDIX 2: R SOFTWARE FILES</td>
<td>79</td>
</tr>
</tbody>
</table>
List of Figures

Figure 1: Levelized cost of electricity (LCOE) for different sources in 2014 and 2025 (projection) .................................................1
Figure 2: Example of wind speed curve of a turbine .......................................................................................................................2
Figure 3: Operator brought on offshore wind platform by helicopter ...............................................................................................3
Figure 4: NWP grid and parameters ..............................................................................................................................................5
Figure 5: Output of a WRF Model .................................................................................................................................................6
Figure 6: Example of cubic regression spline ..................................................................................................................................16
Figure 7: Map of offshore sites ......................................................................................................................................................17
Figure 8: Ijmuiden mast ..............................................................................................................................................................18
Figure 9: Wind maximum and mean speed plot from Fino1, case 1 ...............................................................................................20
Figure 10: Wind maximum and mean speed plot from Fino1, case 2 ............................................................................................21
Figure 11: Maximum wind speed and SPEED100 plot from Fino1, case 1 ....................................................................................22
Figure 12: Maximum wind speed and SPEED100 plot from Fino1, case 2 ....................................................................................22
Figure 13: Comparison of Short Term Models MAE, Fino1 ...........................................................................................................27
Figure 14: Comparison of Short Term Models Climatology Skill Scores, Fino1 ...........................................................................28
Figure 15: Comparison of Short Term Models Persistence Skill Scores, Fino1 ...............................................................................29
Figure 16: Forecasts Persistence 1h ahead, Fino1 .............................................................................................................................30
Figure 17: Forecasts VAR 1h ahead, Fino1 ........................................................................................................................................30
Figure 18: Forecasts Persistence 12h ahead, Fino1 ..........................................................................................................................31
Figure 19: Forecasts VAR 12h ahead, Fino1 ....................................................................................................................................31
Figure 20: Linear and GAM fits of Fino1 training data ....................................................................................................................32
Figure 21: 1st Comparison of one variable-GAM MAE, Fino1 ........................................................................................................33
Figure 22: 2nd Comparison of one variable-GAM MAE, Fino1 .........................................................................................................34
Figure 23: Comparison of two variables-GAM MAE, Fino1 .............................................................................................................35
Figure 24: Comparison of Linear Model and GAM MAE, Fino1 .....................................................................................................36
Figure 25: Comparison of MAE with different update frequency for predictors, Fino1 .................................................................37
Figure 26: Comparison of Long Term Models MAE, Fino1 ...............................................................................................................38
Figure 27: Comparison of RMSE with different update frequency for predictors, Ijmuiden ............................................................40
Figure 28: Comparison of Long Term Models RMSE, Ijmuiden .....................................................................................................40
Figure 29: Forecasts Persistence 12h ahead, Ijmuiden ......................................................................................................................42
Figure 30: Forecasts GAM 12h ahead, Ijmuiden .............................................................................................................................42
Figure 31: Forecasts Persistence 72h ahead, Ijmuiden ......................................................................................................................43
Figure 32: Forecasts GAM 72h ahead, Ijmuiden .............................................................................................................................43
Figure 33: Description of a boxplot ................................................................................................................................................44
Figure 34: Residuals boxplots for Fino1 GAM, 1h-12h horizons ..................................................................................................45
Figure 35: Residuals boxplots for Fino1 GAM, 12h-24h horizons.................................................................45
Figure 36: Histogram of residuals for Fino1 GAM and horizon 1h.................................................................46
Figure 37: Comparison of GAM MAE with different speed limits, Ijmuiden.....................................................47

List of Tables
Table 1: WRF variables with different altitudes.............................................................................................19
Table 2: WRF variables independent of altitude..........................................................................................19
Table 3: Fino1 short-term forecasts values..................................................................................................26
**Abbreviations**

AIC: Akaike’s Information Criterion

ARIMA: Auto-Regressive Integrated Moving Average

BIC: Bayesian Information Criterion

CFD: Computational Fluid Dynamics

EC: Evaluation Criterion

GAM: Generalized Additive Model

MAE: Mean Absolute Error

NA: Not Available

NWP: Numerical Weather Prediction

RMSE: Root-Mean-Square Error

SS: Skill Score

SSE: Sum of Squared Errors

VAR: Vector Auto Regression

WRF: Weather Research and Forecasting
ACKNOWLEDGMENTS

I would like to thank Pierre-Julien Trombe, my supervisor at Vattenfall, for his invaluable help and guidance throughout this study.

Lars Olesen, head of the Operation and Site Technology section, for facilitating, in every possible way, the procedures of my relocation to Denmark.

The Operation and Site Technology team: Anette Hansen, Jens Madsen, Benjamin Martinez, Jesper Nielsen Nissen, Marie Cecilie Pedersen and Anders Sommer, as well as all the Vattenfall employees from the Kolding office, for their very warm welcome.

And Miroslav Petrov, for his constant support in every phase of this KTH degree project.

Kolding
August 2016
1 INTRODUCTION

1.1 Background

Vattenfall is one of Europe’s leading energy companies, whose core activity was construction and exploitation of hydropower dams at the time of its creation. However the company has since extended its activity to the production of electricity and heat by other energy sources, as well as distribution, sales and trading (Vattenfall website page 1, 2016). Vattenfall is currently diversifying its portfolio into renewable energy, notably wind: it is now the 2nd largest operator of offshore wind farms in the world, with 1.3 GW of installed capacity (Vattenfall website page 2, 2016).

![Figure 1: Levelized cost of electricity (LCOE) for different sources in 2014 and 2025 (projection)](costing.irena.org/charts/power-generation-summary-charts.aspx)

At the moment the costs of offshore wind projects remain relatively high compared to traditional fossil fuel energy sources (as shown in Figure 1; offshore wind is the second source from the left). Therefore, a consequence of Vattenfall’s diversification in wind power is that the company’s earnings and costs are becoming more and more dependent on the knowledge of weather conditions: the different reasons for this situation will be detailed in the following paragraphs.
The reduction of installation and construction costs of an offshore project is a priority for Vattenfall. During the installation phase, transportation of the turbines from construction factory to wind site is much more dependent on weather conditions for offshore than for onshore projects, since transport vessels are involved. Because these ships are not operated in case of bad weather for reasons of safety, extreme winds would force to delay the transportation (Flodérus, 2008), which would cause extra transportation costs for Vattenfall. Therefore the ability to predict as early as possible weather windows with low wind speeds would provide an opportunity to reduce the installation costs by shipping the turbines in favorable time periods.

![Wind Speed Curve](https://elements认清energy.blogspot.se)

Figure 2: Example of wind speed curve of a turbine  
*Source: elements认清energy.blogspot.se*

The profitability of an offshore wind farm during the operation phase is also highly dependent on the weather conditions. First because extreme wind speeds will damage turbines if they are kept in operation while experiencing bad weather, which is why most turbines models shut down automatically once a certain speed has been reached: Figure 2 shows a turbine power curve (i.e. the theoretical relationship between wind and electrical power) with a shutdown speed of 25 m/s. But being able to anticipate extreme wind speeds will allow a wind farm operator to optimize the shutdown process, and prevent some of the energy losses that an automatic but abrupt shutdown would have involved.

In addition, maintenance operations at high heights (illustrated by Figure 3) are suspended in case of bad weather (Flodérus, 2008), preventing the maintenance of turbines’ parts. Therefore the turbine’s components lifetime will decrease, resulting in extra costs of new parts’ purchase. Those are additional arguments for Vattenfall to issue weather and wind speed predictions.
Of course, the security of operators must also be guaranteed both during the installation and operation phases. And high wind speeds can pose a great risk to the safety of workers. For all these reasons, accurate forecasting of wind and its associated uncertainty is essential to Vattenfall.

1.2 Objectives

The primary objective of the thesis is to investigate whether the maximum wind speeds at selected offshore sites could be predicted with an important enough accuracy, using well-chosen statistical models. If the experimental results were to confirm that it was indeed possible, the forecasting tools developed within this thesis project could become a part of Vattenfall’s in-house weather forecasting system in the future.

To achieve this aim, the second objective of the thesis is to specify in detail the methodology used for the study and its concrete applications in a written report.

1.3 Methodology

Prior to the development of the forecasting system, several preparatory steps were implemented. First, the data required for the training process of the model was collected. It consisted of observations coming from three different offshore measurement masts located in the North Sea, between the United Kingdom and the Netherlands; and of wind forecasts produced by a weather forecast model. The quality of the data was also checked, which consisted for example in removing dubious values from the sample. Then the statistical models potentially well suited for the dataset were selected, by identifying the trends and patterns in the data, such as linearity or non-linearity. Finally, the implementation of the chosen models was made using the statistical software R, and their performances compared.
2 WIND FORECASTING STATE OF THE ART

In order to provide a general framework to this thesis, the following section will present an overview of the wind forecasting as engineers perform it nowadays, as well as some possible uses for these forecasts.

2.1 Wind forecasting timescales

Before performing any kind of wind speed forecasting, it is important to define precisely the horizon of the prediction, since this will heavily influence the choice of the prediction technique. It is possible to issue forecasts for very different timescales, from a few hours, days or weeks ahead, to several months or even years ahead (Giebel, 2011), however the precision of the forecast can be expected to decrease when the horizons grow higher.

In the course of this study, a clear distinction will be made between short-term and long-term forecasts. Short-term predictions correspond to a few hours ahead (1-6 hours), while long-term forecasts aim at more than 10 hours and up to 2-3 days ahead. These two timescales call for different methods that will be presented and evaluated herein.

2.2 Wind forecasting approaches

Most of the methods used for wind speeds prediction rely primarily on Weather Research and Forecasting (WRF) models. They consist in high-resolution computer models implementing a Numerical Weather Prediction (NWP). Figure 4 illustrates the principle of a NWP: firstly, the Earth is divided into a 3-dimension grid; then, the differential equations linking the physical parameters of each created cell are formulated. Since these equations are nonlinear, they cannot be solved exactly. In order to issue predictions of future atmospheric properties, the equations must be entered in a simulation software: Figure 5 shows an example of a wind speed forecasting map produced by a WRF model.

This approach requires important computational resources to handle the complexity of the models, but it performs well for horizons superior to 6 hours (Giebel, 2011), i.e. long-term forecasting.
Another approach of prediction, which often complements the meteorological simulations, is to use statistical models. The requirement for these kinds of models is that some significant data (such as the wind speed itself) has been measured for a long enough period at the selected wind site. The statistical method is well suited to a few hours ahead forecasts; that’s because WRF models do not perform better than simple statistics on short horizons, while requiring more calculation resources (Genton, 2012).

There are other wind speeds forecasting approaches that won’t be detailed in this thesis. One example is the use of CFD to reconstruct the wind site topography. This method exploits the physical parameters of the location, such as the roughness of the surface, the air density or the information on the atmospheric boundary layer. In general, this purely physical approach performs quite well for long term forecasting (Giebel, 2011).
Figure 5: Output of a WRF Model

Source: Vattenfall Research Forecast
2.3 Wind forecasting applications

Wind forecasts can be used in different ways by electric utilities like Vattenfall, whose activities include the installation and exploitation of wind turbines.

One of the possible applications of forecasting is providing assistance in balancing an electrical grid that integrates wind turbines. Requirements for an electrical source connected to a grid are to contribute to the equilibrium between the demand of electricity and its production and to the control of the grid frequency. But the issue that arises when wind turbines are added to the grid is the variability of the wind resource; unlike a power plant using fossil fuels, the wind power production cannot be easily controlled. This means that the production has to be forecasted in advance, so that the grid balance can be kept more easily and at lesser costs (Genton, 2012). Using wind speed forecasting to predict the electric production from wind turbines can help achieve this goal.

Another application of wind forecasting, of particular interest for utilities involved in offshore wind activities, is to provide information for the choice of the optimal installation time of a turbine. The installation process begins by the production of the wind turbine components onshore, before they are brought to the site by transportation vessels and assembled by cranes. However, the wind and sea conditions affect heavily the shipping of the turbine parts, and can even prevent the transportation to be made on the scheduled day (Uraz, 2011). This will lead to extra shipping costs for the utilities, which is why they have an interest in having as much information as possible on the best weather window for installation (Dewan, 2014). Wind speed forecasts, in particular the average wind speed at a reference height, can then be used to decrease the installation costs of an offshore project (Heggelund, 2013).

These are only two possible direct uses of wind forecasts, but many other applications exist. For example, one can mention assistance to energy trading, environmental impact analyses, better planning of response to extreme weather events and improved strategies for management of greenhouse gases, among many others (NCAR UCAR website, 2016).
3 STATISTICAL METHODS

After describing the background and framework of the study, the mathematical concepts required for solving the problem will now be introduced: they will mainly consist of various statistical equations that can be used to forecast future values of scientific parameters.

3.1 Statistical Modeling Principles

The goal of statistical forecasting is to predict the value of a response variable $Y$ using a vector of explanatory variables $X = (X_1, X_2, \ldots X_p)$ (James, 2015). The fact that predictors can be not only variables other than $Y$, but also past values from $Y$, has to be stressed.

The relationship between $Y$ and $X$ is of the following form:

$$Y = f(X, \theta) + \varepsilon$$

With $\theta = (\theta_1, \theta_2, \ldots \theta_q)$ a vector of parameters to be estimated, $f$ a function chosen to model the relationship between $X$ and $Y$, and $\varepsilon$ an error term with zero mean and independent from the $X$ value.

To issue a prediction of the response variable $\hat{Y}$, an estimation $\hat{\theta}$ of the parameters vector must be produced. Then a result can be obtained (James, 2015):

$$\hat{Y} = f(X, \hat{\theta})$$
3.2 Reference Models

The statistical models described in this section, though relatively simple, can be used as references to evaluate the performance of more complex models.

3.2.1 Climatology

The climatology consists in averaging a certain quantity of historical data monitored until the time \( T \), in order to issue a forecast for times \( T+h \), with \( h \) the horizon of the prediction.

A mean value can be calculated over all the historical data, or only over the latest values, for example the month just before \( T \). In this study the climatology results will be issued using the wind data from the latest week, which leads to the following climatology equation (Madsen, 2005):

\[
\hat{Y}_{T+h|T} = \frac{1}{n} \sum_{t=0}^{n-1} Y_{T-t}
\]

Where \( n = 24 \text{ hours} \times 7 \text{ days} \).

The climatology model is of interest as a reference model that can be used to compare the performance of other models. Whatever the horizon of the prediction might be, from 1h to 24h ahead, it will have almost the same accuracy in forecasting. So, despite giving relatively poor quality predictions for the first horizons, which are easily beaten by other models, the climatology will often perform on the same level, or even better, than those other models for horizons higher than 24h. To be considered for further study, a model should be able to beat the climatology for at least a few horizons; in the opposite case it probably isn’t accurate enough.

3.2.2 Persistence

The persistence, or naïve method, is a model, in which the values at times \( T+h \) are all supposed to be equal to the latest value monitored, at time \( T \) (Hyndman, 2013):

\[
\hat{Y}_{T+h|T} = Y_T
\]

Despite being extremely simple, this model actually performs well for short horizons, up to 4-6 hours, and definitely better than the climatology (as will be demonstrated later). The reason behind this good performance is that the atmospheric phenomena are actually quite slow processes, meaning that the wind speed-related parameters at a site do not vary that much over the scale of a few hours.
3.3 Short-Term Forecasting Models

In this part the models well suited to predictions for short-term horizons, i.e. no more than 24 hours in advance, will be presented.

3.3.1 Drift method

The drift method is a variation on the persistence method where the forecast at \( T+h \) is equal to the value at time \( T \), plus a term that accounts for the variation of this value over time. This term consists of \( h \), the number of hours passed since \( T \), multiplied by the mean change of the considered variable over one hour, calculated on the whole training set (Hyndman, 2013):

\[
\hat{Y}_{T+h|T} = Y_T + \frac{h}{T-1} \sum_{t=2}^{T} (Y_t - Y_{t-1}) = Y_T + h \left( \frac{Y_T - Y_1}{T-1} \right)
\]

3.3.2 Exponential smoothing

The exponential smoothing model objective is to find a compromise between the persistence method, which only considers the latest values from the training sample, and the climatology, which grants the same importance to the recent values and the ones from a distant past.

With exponential smoothing, all values are taken into account, however the most recent values can be given more importance in the calculation than the older ones. This is accomplished through an average whose terms’ weights decrease exponentially the further they are in the past. The equation, with a smoothing parameter \( 0 \leq \theta \leq 1 \), can be written (Hyndman, 2013):

\[
\hat{Y}_{T+h|T} = \theta Y_T + \theta(1-\theta)Y_{T-1} + \theta(1-\theta)^2Y_{T-2} + \cdots
\]

To obtain an adequate value for the smoothing parameter \( \theta \) in the modelization, the R software exponential smoothing function tests several \( \theta \) values and selects the one that minimizes the sum of squared errors over the training period (Hyndman, 2013):

\[
SSE = \sum_{t=1}^{T} (Y_t - \hat{Y}_{t|t-1})^2
\]

3.3.3 ARIMA models

The ARIMA (acronym for Auto-Regressive Integrated Moving Average) model’s characteristic is to take into account the autocorrelations in the training data, which means that the predictors for the answer variable are actually the variable past values.

An ARIMA model, as stated in its name, combines two types of models, the autoregressive models and the moving average models, with an integration process, which in this case is the opposite of a differencing process.
The first order differenced series $\Delta Y_t$ for a variable $Y_t$ is simply defined as the evolution between terms $Y_t$ and $Y_{t-1}$ (Hyndman, 2013):

$$\Delta Y_t = Y_t - Y_{t-1}$$

A second order differenced serie $\Delta^2 Y_t$ would then be:

$$\Delta^2 Y_t = \Delta Y_t - \Delta Y_{t-1} = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2}$$

An autoregressive model for a response variable $Y_t$ can described by an equation of this form, with $c$ a constant term and $\epsilon_t$ a term of error with random variation and zero mean (Hyndman, 2013):

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \epsilon_t$$

$\phi_1, \ldots, \phi_p$ are the model parameters and $p$ is its order. Therefore the predictor vector for the response variable is in fact $X = (Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p})$ and the parameters vector $\theta = (c, \phi_1, \ldots, \phi_p)$.

While an autoregressive model uses the past values of the response variable $Y_t$, a moving average model uses the past values of the prediction error $\epsilon_t$. Here is the equation’s form of such a model, with $c$ a constant (Hyndman, 2013):

$$Y_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

In the same way $\theta_1, \ldots, \theta_q$ are the model parameters and $q$ is its order.

To define the equation of an ARIMA model in a simpler fashion, the backshift notation is introduced (Hyndman, 2013):

$$bY_t = Y_{t-1}$$

It is then possible to write:

$$\Delta Y_t = Y_t - bY_t = (1 - b)Y_t$$

$\Delta^2 Y_t = Y_t - 2bY_t + b^2 Y_t = (1 - b)^2 Y_t$

And in general a difference of order $d$ can be written $(1 - b)^d Y_t$.

With the different notations introduced, it is now possible to define the equation of an ARIMA($p$, $d$, $q$) model for the maximum wind speed:

$$\Delta^d Y_t = c + \phi_1 \Delta^d Y_{t-1} + \cdots + \phi_p \Delta^d Y_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t$$
Which can be re-written:

\[(1 - \phi_1 b - \cdots - \phi_p b^p)(1 - b)^d Y_t = c + (1 + \theta_1 b + \cdots + \theta_q b^q) \varepsilon_t\]

Where d is the degree of differencing, p the order of the autoregressive part and q the order of the moving average part.

Before being able to forecast the response variable with an ARIMA model, it is necessary to obtain the (p,d,q) well suited to the training data, as well as the parameters \(\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q\). When (p,d,q) are known, the other parameters are obtained by software R through the maximum likelihood estimation, which is similar to the minimization of the sum of squared errors: \(SSE = \sum_{t=1}^{T} \varepsilon_t^2\).

As for (p,d,q) selection, one possibility is to test different values and to find the combination that minimize the corrected Akaike’s Information Criterion (AICc) (Hyndman, 2013).

Akaike’s Information Criterion (AIC) is defined as (with \(k=1\) if \(c\neq 0\) and \(k=0\) if \(c=0\)):

\[AIC = -2 \log L + 2(p + q + k + 1)\]

Where L is the likelihood of data, or the probability that the observed data comes from the model, calculated by R.

Then:

\[AICc = AIC + \frac{2(p + q + k + 1)(p + q + k + 2)}{T - p - q - k - 2}\]

The AICc can be obtained by the function Arima in R. The test process of (p,d,q) values can even be automated by using the function auto.arima, however it increases the risk of error. It is better to compare the AICc manually.

### 3.3.4 Linear and quantile regression

ARIMA models are in fact a specific category of linear regression models. A linear regression model uses linear combinations of multiple predictor variables to issue an estimation of the response variable’s values. The model equation is of the following form (Hyndman, 2013):

\[Y_t = \theta_0 + \theta_1 X_{1,t} + \theta_2 X_{2,t} + \cdots + \theta_k X_{k,t} + \varepsilon_t\]

With \(Y_t\) the response variable at instant t, \(X_{1,t}, \ldots, X_{k,t}\) the different predictors at instant t, \(\theta_0\) the intercept term, \(\theta_1, \ldots, \theta_k\) the model coefficients and \(\varepsilon_t\) the prediction error with zero mean.
A regular linear model can be really useful for forecasting in case all related predictor values are known. The challenge when using the linear regression model for forecasting future values is that the predictor variables can’t be used in a direct way: that’s because a model of the form \( Y_t = \theta_0 + \theta_1 X_{1,t} \) does not help in predicting the response variable at future time \( t \) \( Y_t \), since it also requires the knowledge of the predictor at future time \( t \) \( X_t \).

However, it is possible to use the linear regression functionalities of software R by constructing predictors that are actually the response variable values, but shifted several hours in the past. The predictor values at times \( t-1, t-2 \ldots \) are then exploitable to predict the response variables at time \( t \), and the model is actually similar to the auto-regressive model previously defined, with \( X = (Y_{t-1}, Y_{t-2}, \ldots Y_{t-k}) \).

It is also possible to add past values of predictors other than \( Y \) to the model, they are integrated in the equation in the same way, and their coefficient must be determined. These kinds of predictors are referred as exogenous parameters.

Once again, in order to find the appropriate coefficients for the model, the sum of squared errors is minimized on the training period.

As for the quantile regression, it is a model that uses exactly the same equation form as the linear regression, a linear combination of predictors, and it will also be transformed into an auto-regressive model to issue forecasts of maximum wind speed in the future.

However, it differs from the linear regression in the evaluation of the model coefficients: instead of the least square method, which was based on the calculation of mean value of the training maximum wind speed, it is the median value of the training set that will be used for the coefficients’ evaluation. This often allows obtaining a better precision in the forecast than with the linear regression.

### 3.3.5 VAR models

The VAR (acronym for Vector Auto Regression) models allow to consider all the correlations between every variable in the model, whether they consist of autocorrelations or correlations between two different variables. For this reason, a VAR model is expected to better represent the dynamics between the response variable and its predictors.

The two equations of such a model can be written (Hyndman, 2013):

\[
\begin{align*}
Y_{1,t} &= c_1 + \theta_{11,1} Y_{1,t-1} + \theta_{12,1} Y_{2,t-1} + \cdots \\
&\quad + \theta_{11,p} Y_{1,t-p} + \theta_{12,p} Y_{2,t-p} + \varepsilon_{1,t} \\
Y_{2,t} &= c_2 + \theta_{21,1} Y_{1,t-1} + \theta_{22,1} Y_{2,t-1} + \cdots \\
&\quad + \theta_{21,p} Y_{1,t-p} + \theta_{22,p} Y_{2,t-p} + \varepsilon_{2,t}
\end{align*}
\]
Where \( \theta_{11,p} \) is the coefficient accounting for the influence the \( p^{th} \) lag of \( Y_1 \) has on itself, \( \theta_{12,p} \) the coefficient accounting for the influence the \( p^{th} \) lag of \( Y_2 \) has on \( Y_1 \)...

Using matrix notations, the previous equations can be rewritten with any:

\[
Y_t = c + \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \cdots + \theta_k Y_{t-k} + \varepsilon_t
\]

With:

\[
Y_t = \begin{bmatrix} Y_{1,t} \\ Y_{2,t} \\ Y_{3,t} \\ \vdots \end{bmatrix},
\]

\[
c = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \end{bmatrix},
\]

\[
\varepsilon_t = \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \\ \vdots \end{bmatrix},
\]

\[
\theta_i = \begin{bmatrix} \theta_{11,i} & \theta_{12,i} & \theta_{13,i} & \cdots \\ \theta_{21,i} & \theta_{22,i} & \theta_{23,i} & \cdots \\ \theta_{31,i} & \theta_{32,i} & \theta_{33,i} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}
\]

An element of interest is the covariance matrix of the estimated residual error term (here with 2 equations only):

\[
\Sigma = \begin{bmatrix}
\text{cov}(\varepsilon_{1,t}, \varepsilon_{1,t}) & \text{cov}(\varepsilon_{1,t}, \varepsilon_{2,t}) \\
\text{cov}(\varepsilon_{2,t}, \varepsilon_{1,t}) & \text{cov}(\varepsilon_{2,t}, \varepsilon_{2,t})
\end{bmatrix}
\]

The covariance of two variables \( x \) and \( y \) being defined, given their respective expected values \( E(x) \) and \( E(y) \), as:

\[
\text{cov}(x, y) = E[(x - E(x))(y - E(y))]
\]

For the choice of the order \( p \) of the model, the R function VARselect can directly give the optimal order for the training sample, which was found using several evaluation criteria, including the Bayesian Information Criterion (BIC), which is often a robust criterion of choice. It is defined as (Zivot, 2006):

\[
BIC(p) = \ln|\Sigma(p)| + \frac{\ln T}{T} pn^2
\]

The residual covariance matrix can actually be calculated through this equation (Zivot, 2006):

\[
\tilde{\Sigma}(p) = \frac{1}{T} \sum_{t=1}^{T} \tilde{\varepsilon}_t \tilde{\varepsilon}_t'
\]

As for the coefficients of the model, they are once again estimated by the least square method, which means that the coefficients that minimize:

\[
SSE = \sum_{t=1}^{T} \varepsilon_{1,t}^2 = \sum_{t=1}^{T} (Y_{1,t} - c_1 - \theta_{11,1} Y_{1,t-1} - \theta_{12,1} Y_{2,t-1} - \cdots - \theta_{11,p} Y_{1,t-p} - \theta_{12,p} Y_{2,t-p})^2
\]

are chosen, and the same procedure is applied to \( Y_2 \).
3.4 Long-Term Forecasting Models

In this part the focus will be on the models used for maximum wind speed predictions on long term, i.e. up till some days in advance.

3.4.1 Linear Regression

The linear model that was presented in part 3.3 can also be used for long-term predictions. If $Y_t$ is the response variable at instant $t$ and $X_{1,t}, ..., X_{p,t}$ different predictors at instant $t$, then a linear regression model is defined as:

$$ Y_t = \theta_0 + \sum_{i=1}^{p} \theta_i X_{i,t} + \epsilon_t $$

With $\theta_0$ the intercept term, $\theta_1, ..., \theta_p$ the model coefficients and $\epsilon_t$ the prediction error with zero mean. The coefficients of the model $\theta_0, \theta_1, ..., \theta_k$ can be obtained by using software R, which estimates them by minimizing the sum of squared errors: $SSE = \sum_{t=1}^{T} \epsilon_t^2 = \sum_{t=1}^{T} (Y_t - \theta_0 - \theta_1 X_{1,t} - \cdots)^2$ over the training set $(Y_1, ..., Y_T)$.

The main difference with the linear model used for short-term forecasting is that the historical values from the response variable ($Y_{t-1}, Y_{t-2}, ...$) won’t be sufficient for predicting accurately the response variable on the long term: other type of predictors will have to be introduced.

3.4.2 Generalized Additive Models (GAM)

Because linear models are not always able to issue long-term forecasts with a high enough precision, other models have to be introduced. The generalized additive models (or GAM) are more complex to implement, however they can offer better performances, especially when the correlations between the response variable and the predictors are non-linear.

If $Y_t$ is the response variable at instant $t$ and $X_{1,t}, ..., X_{p,t}$ different predictors at instant $t$, then a GAM model can be written as (Hastie 1986):

$$ Y_t = \beta_0 + f_1(X_{1,t}) + f_2(X_{2,t}) + \cdots + f_p(X_{p,t}) + \epsilon_t $$

With $\beta_0$ the intercept term and $\epsilon_t$ the prediction error with zero mean. $f_1, f_2, ..., f_p$ are smooth functions of respectively $X_{1,t}, X_{2,t}, ..., X_{p,t}$. A smooth function is defined mathematically as having continuous derivatives of different orders over its domain; in graphical terms a smooth curve has no rough edges or sharp change of direction.

$f_1, f_2, ..., f_p$ are chosen in order to fit the training data as well as possible, while remaining as smooth as possible. An example is provided on Figure 6: the regression curve does not exactly fit all the points from the data, but doing so would make it very rough-looking, instead of rounded and slowly-fluctuating. In conclusion the functions used in a GAM model are a compromise between precision and smoothness.
Figure 6 illustrates a cubic regression spline, the only types of smooth functions that will be used for this thesis’ previsions. A cubic regression spline $s_k$ can be defined as (Rodriguez, 2001):

$$s_k(x) = \theta_k,0 + \theta_k,1x + \theta_k,2x^2 + \theta_k,3x^3 + \sum_{i=1}^{n} \lambda_k,i(x - \xi_k,i)_+^3$$

With $\xi_{k,1} < \xi_{k,2} < \cdots < \xi_{k,n}$ the knots through which the curve will have to go, $\lambda_{k,1}, \ldots, \lambda_{k,n}$ constants giving control over the degree of the smoothing, and:

$$(x - \xi_{k,i})_+ = \begin{cases} x - \xi_{k,i}, & \text{if } x > \xi_{k,i} \\ 0, & \text{if } x \leq \xi_{k,i} \end{cases}$$

The software R can once again be used to fit a GAM model: the parameters $\theta_{k,0}, \theta_{k,1}, \theta_{k,2}, \theta_{k,3}$ of each cubic regression spline that minimize the sum of squared errors (SSE) will be selected, so that the model fits the data as well as possible. The smoothing constants $\lambda_{k,1}, \ldots, \lambda_{k,n}$ will be chosen by applying the method of generalized cross-validation, which consists in minimizing the following criteria (Aldrin, 2012):

$$-2 \log L \cdot n / (n - p)^2$$

With $L$ the likelihood of data, $n$ the number of observations and $p$ the number of parameters.
4 EXPERIMENTAL DATA

4.1 Presentation of data

In order to study the viability of planned offshore wind projects, Vattenfall often acquires data from measurement masts located at sites that present good wind conditions. Apart from different wind speed measurements (mean, maximum, minimum... as well as direction) obtained by anemometers, other parameters of interest are monitored, such as temperature, air humidity, pressure, precipitation, wave height, captor states... Through the years, the company has accumulated a great quantity of wind speed measurements for different sites, which can be used for statistical studies.

Initially, this thesis project was intended to focus on the measurements from three masts located in the North Sea, between the UK and the Netherlands: Fino1 (in red on Figure 7), Ijmuiden (in green on Figure 7) and East Anglia (in yellow on Figure 7).

A preliminary task had to be performed on the measurement sets from each site. It consisted in transforming the raw information from the sensors into exploitable data, by removing the duplicated values, putting the data in chronological order, attributing the same code to all missing values (which could be marked as NA i.e. Not Available, or as negative values), resampling the data on a 1 hour step.
In the end, the decision was made to focus on sites Fino1 and Ijmuiden for the prediction work, and not to use the data from East Anglia.

In order to be exploitable for forecasting, data sets were divided into a training set and an evaluation set, the evaluation set being more recent than the training set. Both Fino1 and Ijmuiden data, following the resampling process, had a temporal resolution of 1h (interval of 1h between two values of a variable), and the forecasts had a similar resolution. Since the training and evaluation sets had at least 1 year of data each for every model tested, the training and evaluation sets had both contained a minimum of 8760 values.

In addition to the measurement data from the masts, weather forecasts for the wind sites were also used. They were issued by an in-house WRF computer model, which was developed by members of Vattenfall Operation and Site Technology section. Table 1 names and describes variables from this WRF model that are monitored at different altitudes, while Table 2 shows WRF variables independent of altitude.
### Table 1: WRF variables with different altitudes

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Variables at different altitudes (in meters)</th>
</tr>
</thead>
<tbody>
<tr>
<td>U (m/s)</td>
<td>Earth-relative mean wind speed u-component</td>
<td>U20, U50, U100, U150</td>
</tr>
<tr>
<td>V (m/s)</td>
<td>Earth-relative mean wind speed v-component</td>
<td>V20, V50, V100, V150</td>
</tr>
<tr>
<td>SPEED (m/s)</td>
<td>Mean wind speed magnitude</td>
<td>SPEED20, SPEED50, SPEED100, SPEED150</td>
</tr>
<tr>
<td>DIRECTION (degrees)</td>
<td>Earth-relative wind direction</td>
<td>DIRECTION20, DIRECTION50, DIRECTION100, DIRECTION150</td>
</tr>
<tr>
<td>T (°C)</td>
<td>Temperature</td>
<td>T2, T20, T50, T100, T150</td>
</tr>
</tbody>
</table>

### Table 2: WRF variables independent of altitude

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SST (°C)</td>
<td>Sea Surface Temperature</td>
</tr>
<tr>
<td>PBLH (m)</td>
<td>Planetary Boundary Layer Height</td>
</tr>
<tr>
<td>PSFC (Pa)</td>
<td>Pressure at the Surface</td>
</tr>
<tr>
<td>SLP (hPa)</td>
<td>Sea Level Pressure</td>
</tr>
</tbody>
</table>
4.2 Analysis of data

Using the terminology introduced in Part 3 to describe the data from this study, the response variable to be predicted is the maximum wind speed around the altitude of 100m. This height was selected because at the moment, it is often the height of newly constructed commercial wind turbines (WIndustry website, 2016).

Before going any further, it is very important to know what exactly does “maximum wind speed” mean in the frame of this study, since there are so many possible definitions of the concept. In this case it corresponds to the 3-seconds wind gust: the wind speed value is measured every three seconds over an interval of 1 hour, and the maximum of those values is taken.

If some predictor variables $X$ have been measured at an offshore site until an instant $T$, a forecast at a horizon $h$ could be made, giving a prediction of the maximum wind speed at instant $T+h$ (the time step being the hour in this study).

As seen in Chapter 3, good predictors of future maximum wind speeds are obviously past values of maximum wind speed. But statistical models can also include other predictors: for example the past mean wind speeds at the site.

The choice of mean wind speed as a predictor can be justified by the strong correlation between the evolution of mean and max wind speed over time, as can be seen on Figure 9.

![Wind maximum and mean speed plot from Fino1, case 1](image)

However, the correlation between the two variables is not always so important, as Figure 10 shows: the maximum speed peaks much higher than the mean speed around 3800h. Error of prediction will obviously be made, but mean wind speed remains a predictor of interest.
Other possible predictors are the weather forecasts issued for the wind sites. An example of forecast that could be of interest is SPEED100 (see Table 1), because 100m is the height of the maximum wind speed forecast in the case of Fino1. In the same way as the real mean wind speed values, SPEED100 can be highly correlated with the maximum wind speed (as in Figure 11), even though it is not always the case (as Figure 12 demonstrates).

In any case, SPEED100 could potentially be an even better predictor than the mean wind speeds shown on Figures 9 and 10, which are actually unavailable for the time of the maximum wind speed forecast: only past and present values are known and exploitable by the models. On the opposite, SPEED100 is a mean wind speed forecast; therefore the correlation between SPEED100 and the maximum wind speed shown on Figure 11 can be exploited more directly.
Figure 11: Maximum wind speed and SPEED100 plot from Fino1, case 1

Figure 12: Maximum wind speed and SPEED100 plot from Fino1, case 2
5 METHODOLOGY

The experimental process applied in this project will now be described.

5.1 General process

First the statistical models from Part 3 will be formulated and wind speed data and weather forecasts will be used to train each of them, i.e. the equations’ parameters will be estimated using either the maximum likelihood method or the minimization of conditional sum-of-squares method. Then maximum wind speed forecasts will be issued over a posterior dataset and the results of the model compared to the reality, after which the models giving the most precise results will be selected. To give a concrete example, real observations for the whole year 2014 will be used to train the models, whose equations will in turn be used to forecast maximum wind speeds at the site for every hour of 2015. After that, it will be possible to compare the forecasts with the maximum wind speed real observations of 2015.

It is worth noting that the training window (i.e. the amount of training data determined by a beginning date and an end date) does not have to be fixed, with a single evaluation of the model’s parameters: for example the training window could only contain the values of the month anterior to the forecasting date, and the model parameters be updated with every change of date. This can allow for better forecasting precision since, as discussed above, the training wind speeds measured on dates close to the instant of forecasting are often more significant than long past measurements.

The training process will also be slightly different for short-term and long-term forecasting: while the only data used for short-term models will be maximum and mean wind speed, the weather forecasts for long-term forecasts will contain a great number of different predictors. In order to deal with this, a forward selection process will be implemented: the one predictor variable that will offer the best forecast will be determined and added to the tested model. Then each remaining predictor will be added to this model, one at a time, and the predictor improving the most the precision of the forecast will be kept. The process will be repeated until no addition of new variable improves the model.
5.2 MAE, RMSE and Skill Score

The criteria that will serve to evaluate the precision of the forecasts are the mean absolute error (MAE) and the root-mean-square error (RMSE). They are defined by the following equations (Madsen, 2005):

\[
MAE(h) = \frac{1}{n} \sum_{t=T_1}^{T_{end}} |Y_t - \hat{Y}_{t|t-h}|
\]

\[
RMSE(h) = \sqrt{\frac{1}{n} \sum_{t=T_1}^{T_{end}} (Y_t - \hat{Y}_{t|t-h})^2}
\]

Where \(Y_{T_1}, \ldots, Y_{T_{end}}\) are the maximum wind speed values available for the year 2015 and for which predictions were issued, \(n\) the number of those values, \(h\) the forecast horizon of the prediction and \(\hat{Y}_{T_1|t-h}, \ldots, \hat{Y}_{T_{end}|T_{end}-h}\) the corresponding forecasts made \(h\) hours in the past.

Another metric used for the model comparison will be the skill score, which allows to obtain more precise information on how much a statistical model improves the precision of the maximum wind speed forecast compared to another. The skill score equation for horizon \(h\), with \(EC_{ref}\) the evaluation criterion of the reference model and \(EC\) the evaluation criterion of the tested model, is (Madsen, 2005):

\[
SS(h) = \frac{EC_{ref}(h) - EC(h)}{EC_{ref}(h)}
\]

The process called k-fold cross-validation is interesting to mention in this part: it is an often-applied method, which allows verifying that a statistical model will produce accurate forecasts. It consists in dividing the data set into \(k\) subsamples, and for each sample, forecasting its values by using the model fitted with the data outside of the sample. Mean errors of prediction for each subsamples can then be obtained, and finally the mean of the \(k\) previous errors gives information on the forecast quality: minimizing this value by choosing adequate parameters for the model will result in better forecasts (Hyndman, 2010). However, this process cannot be applied for the wind speed forecasting, since the considered data is time-dependent: the most ancient values would be predicted by using a model that is fitted with the most recent values, which is contradictory.
6 RESULTS

In this section the methodology described in Chapter 5 will be applied to the wind data introduced in Chapter 4, thus providing information on the accuracy of maximum wind speed forecasts at offshore sites. The R software files that were produced by the author for this task are presented in detail in Appendix 2.

6.1 Short-Term Forecasting

In this part forecasts up to 24h in advance will be issued and their precision assessed. The statistical models will use historical maximum and mean wind speed values as predictor variables.

6.1.1 Initial Assessment

The short-term models introduced in Chapter 3 are commonly used in prevision, but not all of them were necessarily well suited for maximum wind speed forecasting. For this reason a pre-selection process was conducted: the values of MAE and RMSE were calculated for each model, but only up to a horizon of 3h. The results are summarized in Table 3.
An analysis of those results shows that the persistence has lower MAE and RMSE than every model in the table, except the VAR model. Since persistence is a simpler model than the drift method, exponential smoothing and Arima, but performs better than them, the decision was made to not investigate further their performance, and to focus on VAR model.
6.1.2 Results and Analysis

In order to analyze the quality of their predictions, the MAE and RMSE of the following models were plotted as functions of horizon $h$: climatology, persistence, quantile regression and VAR models with different training windows. Figure 13 shows the MAE plot (with VAR corresponding to a fixed training window of at least 1 year, VAR-m1 to a moving training window of 1 month before the forecast…).

For the sake of clarity, the short-term RMSE plot of Fino1, as well as the short-term MAE and RMSE plots of Ijmuiden, have been placed in Appendix 1.

A first observation is that MAE and RMSE plots have similar shapes, even if the values differ; only one evaluation criterion is enough to compare the performance of the different models in this case. In the same way, the results for site Fino1 do not differ very much from the Ijmuiden results. This is understandable since the offshore sites are located close to each other and are therefore likely to experience similar wind conditions.

![Figure 13: Comparison of Short Term Models MAE, Fino1](image)

As can be seen, climatology has a mean error lying between 4-4.5 m/s for every horizon. The other models have much lower errors for the first horizons (1 m/s only for horizon 1h), but their MAE quickly increase and get closer to the climatology levels for higher horizons (for quantile and VAR models, around 4 m/s for horizon 24h). The climatology even beats the persistence for horizons 21h-24h.
Persistence, quantile and VAR models are very hard to distinguish in terms of performances for the first horizons, however as horizons become higher it is clear that the persistence performs worse than the quantile model, which itself performs slightly worse than the VAR model with fixed training window. As for comparison of VAR models with moving training windows, varying from 1 to 12 months before the date of forecasting, the best performance was obtained with a training window of 1 month, also beating the fixed-window VAR. A similar result is obtained for Ijmuiden.

Additionally, skill scores were calculated for each horizon and site. The evaluation criteria considered were MAE and RMSE, the reference model was climatology then persistence, and the tested models were quantile and several VAR models. The results provide the percentage of improvement of a model compared to “reference” models. Only the skill scores using the MAE of Fino1 are shown here, the plots for Fino1 RMSE and Ijmuiden can be found in annex 1.

Figure 14 shows that quantile and VAR models, compared to climatology, give a forecasting precision improved by more than 70% for horizon 1h. But the improvement percentage quickly falls, and becomes close to 10% for 1-day ahead forecasts. On the contrary, Figure 15 illustrates that quantile and VAR scores, when compared to persistence’s, grow from 2-4% for the first horizons to 8-13% for horizon 24h.

![skill score MAE climatology Fino1](image_url)

Figure 14: Comparison of Short Term Models Climatology Skill Scores, Fino1
6.1.3 Forecasts versus reality

As shown in the previous section, the most precise model for short-term maximum wind speed forecasts at the offshore sites is a VAR model with a sliding 1-month training window. In order to better visualize how VAR_m1 previsions improve on forecast from a simple persistence, the real maximum wind speed values measured at Fino1 were plotted together with the values predicted by persistence and VAR models, for different short horizons.

Figures 16 and 17 show persistence and VAR 1h-ahead forecasts, and both forecasts are very close to the real values. It is also the case for 3h-ahead forecasts (in Annex 1), in accordance with the MAE plot from Figure 14, which showed that persistence and VAR have almost the same level of precision for very short horizons.

On figures 18 and 19 the horizon is 12h: in this case (as for 6h horizon plots in Annex 1) it seems that the VAR forecasts follow more quickly the extreme variations of the response variable than the persistence. This is also in accordance with the MAE plot, which showed lower forecasting error for VAR model at higher horizons.
Figure 16: Forecasts Persistence 1h ahead, Fino1

Figure 17: Forecasts VAR 1h ahead, Fino1
Figure 18: Forecasts Persistence 12h ahead, Fino1

Figure 19: Forecasts VAR 12h ahead, Fino1
6.2 Long-Term Forecasting

In this part forecasts up to 72h in advance will be issued and their precision assessed. The statistical models will use historical maximum wind speed values and forecasts from Vattenfall’s in-house WRF model as predictor variables.

6.2.1 Initial Assessment

In the case of short-term forecasting, an important number of models were selected and tested, but only the ones with the best performance were investigated further. On the contrary, the linear model and GAM were specifically selected for long-term forecasting, based on the results of the study “Probabilistic maximum-value wind prediction for offshore environments”. It is reviewed in the appendices to this report (Staid, 2013), and will be referred to as “the Andrea Staid Study” for the rest of this work.

However, to ensure that both models were suitable for this study, some fitting curves were produced for the training data. For example Figure 20 illustrates that both a linear model and a GAM, using SPEED100 as a predictor, apparently seem to fit the response variable in a correct manner. Therefore the rest of the study was carried out with both models.

![Figure 20: Linear and GAM fits of Fino1 training data](image-url)
6.2.2 Results and Analysis

The forward selection process described in Chapter 5 was applied to a GAM model: the first step was to find the best predictor among the weather forecasts. Figure 21 presents an MAE plot with some examples of the tested predictors for Fino1 GAM. The variables’ descriptions can be found in Table 3 further above.

The RMSE plot, which has a similar shape, can be found in Appendix 1, together with all the other RMSE plots for Fino1 long-term forecasts.

Figure 21 clearly shows that the best precision for the maximum wind speed forecast at 100m is obtained when SPEED100, a forecast of the mean wind speed at 100m height, is used as predictor variable (together with historical maximum wind speed values).

![Figure 21: MAE Comparison of one variable-GAM MAE, Fino1](image_url)
The forecasts using SPEED100 as predictor were then compared to forecasts using SPEED50 and SPEED150. Figure 22 shows that they performed very well too, so they could have been selected as the first variable in the forward selection.

However, since the response variable was the maximum wind speed at 100m, SPEED100 was chosen.

Figure 22: 2nd Comparison of one variable-GAM MAE, Fino1
The following step in the forward selection was to test whether adding another variable to the GAM model with SPEED100 as predictor would improve the precision of the forecasts. Figure 23 compares MAE when no new variable is added to the model, and when SPEED50 and T100 are added: it can be seen that those variables do not improve the forecasts with predictor SPEED100, or only very slightly.

Additions of other weather forecasts were also tested, with the same results. It was therefore decided to keep only SPEED100 as a predictor of maximum wind speed, together with past maximum wind speeds.

Figure 23: Comparison of two variables-GAM MAE, Fino1
The performance of linear models for long-term forecasting was also assessed. Figure 24 compares the MAE of a GAM and a linear model, both using SPEED100 and past maximum wind speed values as predictors. The GAM performs slightly better than the linear model, except for the horizons 40-50h (which is even more the case for the RMSE plot, presented in detail in Appendix 1).

In the remainder of the study only the GAM will be studied, however the linear model would also have been an excellent choice.

![Figure 24: Comparison of Linear Model and GAM MAE, Fino1](image)

Figure 24: Comparison of Linear Model and GAM MAE, Fino1
All the forecast results from the GAM presented until now used weather forecasts that were updated every hour, offering a great quantity of data, but also requiring important computational resources. In order to determine whether changing the update frequency of WRF predictors would be profitable, forecasts were issued with a GAM using SPEED100 values updated only once a day. Figure 25 compares MAE for both cases (the legend for updates once a day is GAM simple SPEED100).

The mean error of the GAM with updates every hour grows from 1 m/s for horizon 1h to approximately 1.7 m/s for horizon 10h, and remains stabilized at this value for higher horizons. The MAE of GAM updated once a day follows a similar pattern, but has much more oscillations around the 1.7 m/s value, making it less predictable. So the decision was taken to keep updating the weather forecasts every hour.

Figure 25 also shows the precision of a forecast that only uses SPEED100 values to directly predict the maximum wind speed, without any use of a statistical model (the legend is WRF SPEED100). It gives an MAE of approximately 2.7 m/s, more than the MAE of the two GAM. It makes sense that using a raw mean wind speed, SPEED100, to predict a maximum wind speed doesn’t make for very accurate forecasts. Therefore, this confirms the benefit of using statistical models for maximum wind speed forecasting.

Figure 25: Comparison of MAE with different update frequency for predictors, Fino1
Finally, the GAM with SPEED100 as predictor is compared to the reference models: the persistence and climatology. Figure 26 shows that the GAM MAE, going from 1 m/s for horizon 1h to a stabilized value 1.7 m/s for horizons higher than 10h, performs better than the climatology for every horizon (its MAE is comprised between 4 and 5 m/s). As for the persistence error, it has the same level as the GAM error for horizons 1h and 2h, but it grows quickly, reaching 6 m/s for horizon 72h. The skill score plots (in Appendix 1) quantify the improvement percentages of GAM as 60-65% compared to climatology for high horizons, and as high as 70% compared to persistence for horizon 72h.

![Figure 26: Comparison of Long Term Models MAE, Fino1](image)

-38-
As for the Ijmuiden site, a forward selection process was also applied to select the best predictor variables for the forecasts of maximum wind speed, this time at a height of 85m. The different plots illustrating the complete process can be found in Appendix 1.

As for the results, the best forecasting precision was obtained with a GAM having SPEED50 and SPEED100 as predictor variables. The response variable is the maximum wind speed at 85m, so it makes sense that predictions of mean wind speeds at 50m and 100m, the two closest available altitudes for weather forecasts, are the best predictors.

A linear model with SPEED50 and SPEED100 as predictor variables was also tested. Its precision was close to the precision of the GAM. However the GAM performed slightly better for horizons higher than 10-20h (as seen on the plot in Annex 1), for this reason it was chosen over the linear model.

In addition, maximum wind speed forecasts were made using values of SPEED50 and SPEED100 updated only once a day. Figure 27 shows that the RMSE with predictors updated every hour grows from 1.3 m/s for horizon 1h to slightly more than 2 m/s for horizon 10h, and remains stabilized at this value for higher horizons. The RMSE of the GAM updated only once a day is however much more unpredictable, sometimes higher sometimes lower than the other GAM. In the end, as for Fino1, the predictor variables were kept updated every hour.

Figure 27 also illustrates the precision of forecasts directly using SPEED50 and SPEED100 values to predict the maximum wind speed, without any statistical model. The RMSE are 3.3 with SPEED50 and 2.8 with SPEED100, more than the RMSE of the GAM. So as is the case with Fino1, it is interesting to use statistical models with predictor variables to issue forecasts for Ijmuiden, and not only the predictors without any statistical treatment.

The GAM with SPEED50 and SPEED100 as predictors was compared to the persistence and climatology models, and performed very well. Figures 35, 36 and 37 further below allow quantifying the improvements in the precision of maximum wind speed forecast that are brought by GAMs.

Finally, the GAM with SPEED50 and SPEED100 as predictors is compared to reference models persistence and climatology. Figure 28 shows that the GAM RMSE, going from 1.3 m/s for horizon 1h to a stabilized value 2.1 m/s for horizons higher than 10h, performs better than the climatology for every horizon (its MAE remains around 5 m/s). As for the persistence error, it has the same level as the GAM error for horizons 1h and 2h, but it grows quickly, becoming superior to 6 m/s for horizons 40h-50h. The skill score plots (in Appendix 1) quantify the improvement percentages of GAM as approximately 60% compared to climatology for high horizons, and close to 70% compared to persistence for horizon 72h.
Figure 27: Comparison of RMSE with different update frequency for predictors, Ijmuiden

Figure 28: Comparison of Long Term Models RMSE, Ijmuiden
6.2.3 Forecasts versus reality

The model selected for long-term forecasting of maximum wind speed values at the offshore sites was GAM, and the analysis of the previous results suggested that the best predictor variables were one or two SPEED parameters, at altitudes close to the altitude of the studied maximum wind speed.

The following plots provide some illustrations of the model performances, by comparing the real maximum wind speed values measured at Ijmuiden with the values predicted by GAM with predictors SPEED50 and SPEED100, and a persistence model, for horizons 12h and 72h (1h and 24h are in Appendix 1).

In accordance with the error plot from the previous part, forecasts from persistence are shown to become less and less precise as horizons grow, from some quite imprecise predictions 12h ahead to the unusable 72h-ahead forecasts. On the contrary, GAM forecasts seem to remain close to the actual maximum wind speed values whatever the forecast horizon is, which was illustrated by the stabilization of the mean error value for horizons higher than 10h.
Figure 29: Forecasts Persistence 12h ahead, Ijmuiden

Figure 30: Forecasts GAM 12h ahead, Ijmuiden
Figure 31: Forecasts Persistence 72h ahead, Ijmuiden

Figure 32: Forecasts GAM 72h ahead, Ijmuiden
6.3 Maximum Forecasting Error

The concepts used in Sections 6.1 and 6.2 to study the errors in the forecasting of maximum wind speeds were the MAE and the RMSE. Both provide information on the mean prediction error of all forecasts issued for a given horizon.

However, knowing how high the forecasting errors will be on average is not enough to exploit the maximum wind speed forecasts for the applications described in the introduction (selection of good weather periods, shutdowns anticipation, improvement of the operators’ security…). It is equally important to be informed of the maximum possible errors for each horizon of prediction, so that the “worst-case scenario”, with a forecast very different from the real value, can be taken into account.

To that aim, the distribution of residuals for the maximum wind speed at Fino1 and Ijmuiden was studied. A residual $res$ is defined as the difference between the real observed value of a response variable, $Y$, and its estimated value obtained through a statistical model, $\hat{Y}$:

$$res = Y - \hat{Y}$$

A really useful tool for studying the distribution of residuals over a dataset is the boxplot. It allows to identify at a single glance the important statistics of a dataset: median, quartiles, maximum and minimum. Figure 33 sums up the different statistics and how they are presented by a boxplot.

![Description of a boxplot](Source: flowingdata.com/2008/02/15/how-to-read-and-use-a-box-and-whisker-plot/)
The boxplots of residuals were drawn for the different sites and models, and horizons of prediction between 1h and 72h. Figures 34 and 35 illustrate that, while the residuals upper quartile remain stabilized at 2 for horizons higher than 4-5h, the “theoretical” maximum is close to 5, with outliers reaching values as high as 15. This means that very high errors of prediction can potentially be made (some Ijmuiden outliers were shown to be higher than 20).

Figure 34: Residuals boxplots for Fino1 GAM, 1h-12h horizons

Figure 35: Residuals boxplots for Fino1 GAM, 12h-24h horizons
However, a more detailed study of the residuals distribution revealed that, while there were indeed some extreme values, they were only a minority in the studied datasets. Figure 36 shows an histogram of residuals, indicating the percentage of residuals between intervals 0-1, 1-2, … In the case of Fino1 GAM, approximately 89% of residuals lie between -2 and 2, and an overwhelming majority of residuals lie between -5 and 5. Other horizons and Ijmuiden were also shown to have around 90% of residuals comprised between -2 and 2 and almost negligible percentage of residuals higher than 5 (or lower than -5).

The conclusion from this analysis is that, while the forecasting errors very rarely exceed 5 m/s, a few isolated residual values can get as high as 15 or 20. However, it must be stressed that the wind speed can never be predicted with a 100% precision: the higher the wind speed the more difficult it is to model accurately, and GAM are not necessarily the best-suited models for forecasting extreme wind events.

To illustrate this situation, the GAM forecasts datasets of Fino1 and Ijmuiden were divided according to the values of real maximum wind speed: MAE and RMSE were calculated for measured maximum speed higher than 15 m/s, and then for measured maximum speed lower than 15 m/s. The value of 15 m/s was chosen because it is generally the limit between “normal” and “strong” winds that is considered in the wind turbine industry (TU Delft website, 2016).
Figure 37: Comparison of GAM MAE with different speed limits, Ijmuiden

Figure 37 shows that the average prediction error for wind speeds lower than 15 m/s, is equal to 1.5 for horizons higher than 10 h, which is slightly lower than the MAE that considers every wind speeds. But the MAE for wind speeds higher than 15 m/s is bigger than 2 except for very short horizons, and whereas the other two MAE remain stable, MAE for high wind speeds grows with the horizons of prediction.

Despite the fact that higher wind speeds are less accurately predicted by GAM than lower ones, the performance of GAM for forecasting maximum wind speeds is satisfying, on average. Nevertheless, it is important to determine the maximum possible error, particularly in case of high wind speed conditions, so that extreme wind events can be better anticipated.
DISCUSSION AND CONCLUSION

Forecasts of maximum wind speeds at offshore sites are of high interest for energy companies like Vattenfall. They can provide great assistance in the reduction of offshore wind farms installation costs, in the optimization of wind parks management, and in ensuring maximal security for operators during the installation and maintenance phases.

The results obtained during the work on this thesis project show that it is possible to forecast maximum wind speeds up to 72h in advance, while maintaining average errors of prediction lower than 2 m/s for every horizon. This level of error is sufficient to make the maximum wind speed forecasts exploitable by wind professionals: operators on wind sites can use the short-term VAR forecasts to anticipate future shutdowns, and long-term forecasts issued by linear models or GAM provide information on the appropriate times for maintenance operations. The outcomes of this work also confirm the results described in the Andrea Staid Study (Staid, 2013), where the forecasting errors are of the same order as those obtained for sites Fino1 and Ijmuiden.

Another positive result is that satisfying forecasts were obtained through the use of statistical models that are relatively simple to describe and implement. Models more elaborate than VAR, linear model or GAM exist, and could be expected to deliver better forecasts since they better represent the complexity of wind, but they weren’t required here. In addition, these models didn’t require a lot of predictor variables to perform well: some historical maximum and mean wind speeds, and one or two forecasts from Vattenfall WRF model, chosen at the right altitudes, were enough to obtain low error scores.

So because the forecasting method developed in this thesis is (relatively) simple and generic, it will be easy to adapt it, so that the methodology can be applied in different conditions. Even if it was designed for offshore wind sites, the process can perfectly be used to predict maximum wind speed on onshore sites. The forecasts were also made for a 72h horizon at most, because this was the forecasting upper limit of Vattenfall WRF model. However, the methodology could just as well be replicated for 7-10 days-ahead forecasts, if the model that delivers predictor variables for the GAM allows it.

There is also some work left to do in order to improve the results from the thesis. First, after obtaining the average GAM error in this study, a probabilistic approach should be implemented in order to obtain the prediction intervals of the forecasted values. This would allow having exhaustive information on the confidence that can be put into the maximum wind speed forecasts. The Andrea Staid Study (Staid, 2013) describes the calculation process of those intervals, through the use of simple linear models. Once the prediction intervals are obtained, the results of the study could be integrated in Vattenfall’s in-house weather forecasting system.

It was shown in section 6.3 that GAMs are not necessarily adapted to the modeling of extreme wind events: other statistical models able to handle those phenomena could be investigated to handle this problem. Furthermore, a moving training window was tested for short-term forecasts, but not for long-term. This was due to the results from the Andrea Staid Study (Staid, 2013), which suggested that a static training window offered the same level of precision as a moving one, provided that the window contained more than 200 observations; the static training windows used for GAM had more than 1000 observations at least. Nevertheless, a moving training window for long-term forecasts should be tested to ensure that it does not improve the precision any further.
For future work, it would be important to implement the methodology on offshore wind sites that are not only limited to the North Sea: it is possible that the models and predictor variables described in the thesis are less adapted to wind conditions different from those of Fino1 and Ijmuiden. Additionally, a method for forecasting wind gusts has recently been developed by Vattenfall (Hawkins, 2016). Forecasting wind gusts could help to better model extreme wind speed events: the possibility to incorporate the gust predictions as predictor of the GAM, like the WRF predictions, should be studied.
REFERENCES


http://orbit.dtu.dk/fedora/objects/orbit:83397/datastreams/file_5277161/content


http://www.norcowe.no/doc/konferanser/2013/presentasjoner%20owo_smi/4_Heggelund_CMR_OW
O%20SMI%202013.pdf

Hyndman, Rob J. (2010). Why every statistician should know about cross-validation. Hyndsite, Wordpress.
http://robjhyndman.com/hyndsight/crossvalidation/

Hyndman, Rob J; Athanasopoulos, George. (2013). Forecasting: principles and practice. OTexts: October
17, 2013.
https://www.otexts.org/fpp

James, Gareth; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert. (2015). An introduction to statistical


Webpage references

Vattenfall website page 1 (last visit August the 11th 2016).
https://corporate.vattenfall.com/about-vattenfall/operations/

Vattenfall website page 2 (last visit August the 11th 2016).

TU Delft website (last visit August the 18th 2016).
http://mstudioblackboard.tudelft.nl/duwind/Wind%20energy%20online%20reader/Static_pages/power_control.htm

NCAR UCAR website (last visit August the 15th 2016).
https://www.ral.ucar.edu/projects/forecasting-for-wind-energy

Windustry website (last visit August the 16th 2016).
http://www.windustry.org/how_big_are_wind_turbines
APPENDIX 1: ADDITIONAL RESULTS PLOTS

Short Term MAE and RMSE plots
Short Term Skill Score plots
skill score MAE climatology ljmuiden

skill score RMSE climatology ljmuiden
Comparison between real values and short term forecasting

Fino1 Forecasts 3h ahead

Wind Speed (m/s)

- Real Max Values
- Persistence Forecasts

Fino1 Forecasts 3h ahead

Wind Speed (m/s)

- Real Max Values
- VAR_m1 Forecasts

time (h)

-59-
Fino1 Forecasts 6h ahead

Wind Speed (m/s)

7000 7050 7100 7150

time (h)

Real Max Values
Persistence Forecasts

Fino1 Forecasts 6h ahead

Wind Speed (m/s)

7000 7050 7100 7150

time (h)

Real Max Values
VAR_m1 Forecasts
Fino1 Long Term RMSE plots
Fino1 Long Term Skill Score plots

**skill score MAE climatology Fino1**

![Graph showing skill score MAE climatology for Fino1 over hours ahead.]

**skill score MAE persistence Fino1**

![Graph showing skill score MAE persistence for Fino1 over hours ahead.]

-64-
Ijmuiden Long Term MAE plots

![Graph 1: MAE Ijmuiden](image1)

![Graph 2: MAE Ijmuiden](image2)
MAE Ijmuiden

![Graph showing MAE (m/s) over hours ahead (h) for different models: climatology, persistence, and GAM SPEED50, SPEED100.](chart)

- Clinatology
- Persistence
- GAM SPEED50, SPEED100
Ijmuiden Long Term RMSE plots

RMSE Ijmuiden

RMSE Ijmuiden

-70-
RMSE Ljuiden

RMSE (m/s)

hours ahead (h)

GAM SPEED50 T50
GAM SPEED50
GAM SPEED50 SPEED100

RMSE Ljuiden

RMSE (m/s)

hours ahead (h)

GAM SPEED50 SPEED100 U50
GAM SPEED50 SPEED100 V100
GAM SPEED50 SPEED100
Ijmuiden Long Term Skill Score plots
Comparison between real values and long term forecasting

Ijmuiden Forecasts 1h ahead

Wind Speed (m/s)

- Real Max Values
- Persist Forecasts

Ijmuiden Forecasts 1h ahead

Wind Speed (m/s)

- Real Max Values
- GAM S50 S100

-75-
Plots of MAE and RMSE with division in low and high wind speeds

![Plots of MAE and RMSE with division in low and high wind speeds](image-url)
APPENDIX 2: R SOFTWARE FILES

prepare_data_ijmuiden_mast_sampled.R

# Script for preparing sampled short term dataset
load("C:/Users/hqv95/Desktop/Data/Wind/Ijmuiden/Ijmuiden_mast_simplified.RData")

# Selection of dates
StartDate <- "2013-01-01 00:00:00"
IdxStartDate <- grep(StartDate,dat[,1])
selectdat <- dat[-(1:(IdxStartDate-1)),]
selectdat[,1] <- sub("\", ", selectdat[,1])
selectdat[,1] <- as.POSIXct(selectdat[,1], format = "%Y-%m-%d %H:%M:%S", tz = "UTC")
rownames(selectdat) <- NULL

# Selection of data type (max and mean speed at 85m)
TypeIdx <- c(1, grep("Wind_Speed_85m_max",names(selectdat)), grep("Wind_Speed_85m_mean", names(selectdat)))
selectdat <- selectdat[TypeIdx]

# Sampling: transformation from values every 10 minutes to values every hour
l_sampled <- dim(selectdat)[1]/6
n_sampled <- 0:(l_sampled-1)
newdat <- data.frame(datetime = as.POSIXct(character(), format = "%Y-%m-%d %H:%M:%S", tz = "UTC"),
ws_max = double(),
ws_mean = double(),
stringsAsFactors = FALSE)
for (k in n_sampled) {
newdat[k+1,1] <- selectdat[6*k+1,1] # Date and Time selection: every hour instead of every 10 min
val_max <- vector(mode = "numeric", length = 6)

# Selection of max wind speed data: 6 values for each hour
for (i in (1:6)) {
  if (!is.na(selectdat[6*k+i,2])) {
    val_max[i] <- selectdat[6*k+i,2] #
  }
}

idx_max <- which(!val_max==0)
val_max <- vector(mode = "numeric", length = 6)

n_max <- length(idx_max)

if (n_max < 3) { # if less than 3 known values
  newdat[k+1,2] <- NA # unknown value for the hour
} else {
  newdat[k+1,2] <- max(val_max) # else value for the hour is the max of 6 (or less) values
}

val_mean <- vector(mode = "numeric", length = 6)

# Selection of mean wind speed data: 6 values for each hour
for (i in (1:6)) {
  if (!is.na(selectdat[6*k+i,3])) {
    val_mean[i] <- selectdat[6*k+i,3]
  }
}

idx_mean <- which(!val_mean==0)
val_mean <- vector(mode = "numeric", length = 6)

n_mean <- length(idx_mean)

if (n_mean < 3) { # if less than 3 known values
  newdat[k+1,3] <- NA # unknown value for the hour
} else {
  newdat[k+1,3] <- sum(val_mean[idx_mean])/n_mean
  # else value for the hour is the mean of 6 (or less) values
}

save("newdat", file = "C:/Users/hqv95/Desktop/Data/Wind/Ijmuiden/Model/Ijmuiden_1hr.RData")
# Forecast with VAR model and rolling training window

```r
load("C:/Users/hqv95/Desktop/Data/Wind/Fino1/Competition3/fino1_1hr_comp3.RData")
```

```r
library(zoo, lib.loc="C:/Users/hqv95/Desktop/R_Packages/R_3_1_1")
library(tsDyn, lib.loc="C:/Users/hqv95/Desktop/R_Packages/R_3_1_1")
library(lmtest, lib.loc="C:/Users/hqv95/Desktop/R_Packages/R_3_1_1")
library(urca, lib.loc="C:/Users/hqv95/Desktop/R_Packages/R_3_1_1")
library(sandwich, lib.loc="C:/Users/hqv95/Desktop/R_Packages/R_3_1_1")
library(strucchange, lib.loc="C:/Users/hqv95/Desktop/R_Packages/R_3_1_1")
library(vars, lib.loc="C:/Users/hqv95/Desktop/R_Packages/R_3_1_1")
```

```r
window <- readline(prompt = "Number of months for the windows : ")
window <- as.numeric(unlist(strsplit(window," ")))
```

```r
l_window <- length(window)
nwindow <- 1:l_window
```

```r
# Separation date between training and prediction set: initial values
end_train_ini <- grep("2015-01-01 00:00:00",newdat[,1])
eval <- newdat[1:end_train_ini,]
l_eval <- dim(eval)[1]
```

```r
dat_pred <- na.approx(newdat[,2:3], na.rm = FALSE)
```

```r
l_pred <- 24
npred <- 1:l_pred
```

```r
month <- 30
hour_month <- month*l_pred
```

```r
nk <- 0:(l_eval-1)
```
# loop
for (t in nwindow) { # loop for every window size (1 to 12 months)

    m_window <- window[t]
    end_train <- end_train_ini

    results <- data.frame(datetime = as.POSIXct(character(), format = "%Y-%m-%d %H:%M:%S", tz = "UTC"),
                      obs = double(),
                      fct_pt = double(),
                      h = integer(),
                      stringsAsFactors = FALSE)

    lag <- data.frame(beginning = as.POSIXct(character(), format = "%Y-%m-%d %H:%M:%S", tz = "UTC"),
                       end = as.POSIXct(character(), format = "%Y-%m-%d %H:%M:%S", tz = "UTC"),
                       lag = double(),
                       stringsAsFactors = FALSE)

for (k in nk) { # loop for every hour

    train <- dat_pred[((end_train-m_window*hour_month):end_train),]
    criteria <- VARselect(train, lag.max=10, type="const")$selection
    n_lag <- criteria["SC(n)"] # find optimal number of lags for every training window
    VAR_100_model <- lineVar(train, lag = n_lag)
    Pred <- predict(VAR_100_model, n.ahead = l_pred)

    lag[k+1,1] <- newdat[end_train-m_window*hour_month, 1]
    lag[k+1,2] <- newdat[end_train, 1]
    lag[k+1,3] <- n_lag # save number of lags for every training window

for (i in npred) { # loop for every prediction horizon
    results[l_pred*k+i,1] <- eval[k+1,1]
    results[l_pred*k+i,2] <- eval[k+1,2]
    results[l_pred*k+i,4] <- i
}
results[l_pred*(k+i-1)+i,3] <- Pred[i,1]

end_train <- end_train + 1

# delete ending lines
results <- results[-((dim(results)[1]-(l_pred^2-l_pred-1)):dim(results)[1]),]

nNA <- 1:(l_pred - 2)
Vec <- 1:(l_pred*(l_pred - 1))
rem <- 1

for (i in nNA) {
  k <- 1
  while (k <= i + 1) {
    rem <- c(rem, l_pred*i + k)
    k <- k + 1
  }
}

VNA <- setdiff(Vec,rem)

for (i in VNA) {
  results[i,3] <- NA
}

# delete useless results
lengres <- 1:dim(results)[1]

for (i in lengres) {
  if (is.na(results[i,2])) {
    results[i,3] <- NA
  }
}
table <- sprintf("C:/Users/hqv95/Desktop/Data/Wind/Fino1/Competition3/fino1_model_3_VAR_window_m\%
s.txt", m_window)
write.table(results, file = table, sep = ",", row.names = FALSE, quote = 1)

RData <- sprintf("C:/Users/hqv95/Desktop/Data/Wind/Fino1/Competition3/fino1_model_3_VAR_window_m\%
%s.RData", m_window)
save("results", file = RData)

table_lag <- sprintf("C:/Users/hqv95/Desktop/Data/Wind/Fino1/Competition3/fino1_model_3_VAR_lag_m\%
t", m_window)
write.table(lag, file = table_lag, sep = ",", row.names = FALSE, quote = 1)

RData_lag <- sprintf("C:/Users/hqv95/Desktop/Data/Wind/Fino1/Competition3/fino1_model_3_VAR_lag_m\%
Data", m_window)
save("lag", file = RData_lag)
}
# Script for preparing long term forecast data set

setwd("C:/Users/hqv95/Desktop/Data/Wind/Fino1/Long Term/")

dat <- read.csv("wrf_fino1.csv")

# Remove NA
idx_NA <- which(is.na(dat[,1]))
dat <- dat[-idx_NA,]

# Remove uncorrect values
idx_rem <- grep("2015-12-11",dat[,1])
dat <- dat[-idx_rem,]

library(lubridate, lib.loc="C:/Users/hqv95/Desktop/R_Packages/R_3_1_1")

dat$init_time <- as.POSIXct(dat$init_time, format = "%Y-%m-%d",
tz = "UTC")
dat$valid_time <- dat$init_time + lubridate::hours(dat$lead_time)

# Forecasts are available with a delay of approx. 8 hours
dat$available_time <- dat$init_time + lubridate::hours(8)

# Drop observations that correspond to the warm-up period of the weather model
idx_drop <- which(dat$valid_time < dat$available_time)
dat <- dat[-idx_drop,]

# For a given datetime, extract the series of forecasts
extract_forecasts <- function(origin_time, lead_time, df) {
  # origin_time is a POSIXct object
  # lead_time is a series of integer
  # df is a data frame containing the forecasts

  dfnew <- data.frame(valid_time = origin_time + lubridate::hours(lead_time))
dfnew <- merge(dfnew, df, by = "valid_time", all.x = TRUE, all.y = FALSE)
# reorder forecasts w.r.t ascending lead time
dfnew <- dfnew[order(dfnew$valid_time, dfnew$lead_time),]
# drop oldest forecasts (=those with the largest lead times)
dfnew <- dfnew[!duplicated(dfnew$valid_time),]
# adjust lead_time
dfnew$lead_time <- lead_time
return(dfnew)

# Model updated every hour
ortime <- seq(ymd("2014-01-01"), ymd("2016-01-01"), by = '1 hour')
ltime <- 1:72

# Construct data frame with historical values from response variable
buildDF <- function(obs, lags, h) {
  df <- data.frame(y = obs)
  lagval <- lapply(lags, function(x, obs, h)  c(rep(NA, h-1+x), obs[1:(length(obs)-h+1-x)]) ,
        h = h, obs = obs)
  df <- cbind(df, as.data.frame(lagval))
  names(df) <- c("Y_t_h", "Y_t", paste("Y_t_m_", lags[-length(lags)], sep = ""))
  return(df)
}

load("fino1_1hr.RData")
dat_val <- newdat
lag <- 5 # 5 lags for instant t+h: t, t-1, t-2, t-3, t-4
idx_test <- lag:dim(dat_val)[1]

for (k in ltime) { # Loop for every horizon
  # Prepare weather forecasts
  dat_me <- extract_forecasts(ortime, ltime[k], dat)
  names(dat_me)[names(dat_me)=="valid_time"] <- "valid_time_t_h"
  dat_me <- dat_me[, !names(dat_me) %in% c("init_time","available_time"))]
  # suppress useless columns

  # Prepare maximum wind speed data
  val <- buildDF(dat_val$ws_max[idx_test[1] - lag:1,idx_test], lags = 1:lag, h = ltime[k])
  dat_st <- cbind(valid_time_t_h=dat_val[1], val)
# Merge weather forecasts and maximum wind speed data by common time
newdat <- merge(dat_st, dat_me, by="valid_time_t_h", all.x=TRUE)
rownames(newdat) <- NULL
names(newdat)[names(newdat)=="lead_time"] <- "lead_time_h"
names(newdat)[names(newdat)=="ws_max"] <- "Y_t_h"

RData <- sprintf("fino1_long_term/fino1_long_term_%s.RData", ltime[k])
save("newdat", file = RData)
}
# Forecast with persistence

```r
library(timeDate, lib.loc="C:/Users/hqv95/Desktop/R_Packages/R_3_1_1")
library(zoo, lib.loc="C:/Users/hqv95/Desktop/R_Packages/R_3_1_1")
library(forecast, lib.loc="C:/Users/hqv95/Desktop/R_Packages/R_3_1_1")

setwd("C:/Users/hqv95/Desktop/Data/Wind/Fino1/Long Term/")

horizon <- 1:72
l_horizon <- length(horizon)

for (k in 1:l_horizon) { # Loop for every horizon
  RData <- sprintf("fino1_long_term/fino1_long_term_%s.RData", horizon[k])
  load(RData)

  # Separation date between training and prediction set
  idx_train <- grep("2015-01-01 00:00:00", newdat[,1])
  eval <- newdat[-(1:idx_train),]
  results <- eval[,c("valid_time_t_h","lead_time_h","Y_t_h","Y_t")]
  names(results)[names(results) == "Y_t"] <- "Y_t_h_pred" # apply persistence
  results[,"Y_t_h_pred"] <- na.approx(results[,"Y_t_h_pred"])
  l_results <- dim(results)[1]

  for (i in 1:l_results) {
    if (is.na(results[i,"Y_t_h"])) {
      results[i,"Y_t_h_pred"] <- NA
    }
  }

  table <- sprintf("Persistence/fino1_persistence_%s.txt", horizon[k])
  write.table(results, file = table, sep = ",", row.names = FALSE, quote = 1)

  RData <- sprintf("Persistence/fino1_persistence_%s.RData", horizon[k])
  save("results", file = RData)
}
```
## Model_climatology_long_term.R

# Forecast with climatology
setwd("C:/Users/hqv95/Desktop/Data/Wind/Fino1/Long Term/")

horizon <- 1:72
l_horizon <- length(horizon)

hour <- 24
week <- 7
hour_week <- week*hour

for (k in 1:l_horizon) {  # Loop for every horizon
  RData <- sprintf("fino1_long_term/fino1_long_term_%s.RData", horizon[k])
  load(RData)

  # Separation date between training and prediction set
  idx_train <- grep("2015-01-01 00:00:00", newdat[,1])
  eval <- newdat[-(1:idx_train),]

  results <- eval[,c("valid_time_t_h","lead_time_h","Y_t_h")]
  results["Y_t_h_pred"] <- NA
  l_results <- dim(results)[1]
  vec <- horizon[k]:l_results
  idx <- idx_train

  for (i in vec) {  # Loop for every date/time
    WS_100_max_train <- newdat[(idx-hour_week+1):idx,"Y_t_h"]  # values from last week
    results[i,"Y_t_h_pred"] <- mean(WS_100_max_train, na.rm = TRUE)  # climatology
    idx <- idx + 1
  }

  table <- sprintf("Climatology/fino1_climatology_%s.txt", horizon[k])
  write.table(results, file = table, sep = ",", row.names = FALSE, quote = 1)

  RData <- sprintf("Climatology/fino1_climatology_%s.RData", horizon[k])
  save("results", file = RData)
}

"
# Model_GAM_SPEED100.R

# Forecast with GAM model

library(mgcv, lib.loc="C:/Users/hqv95/Desktop/R_Packages/R_3_1_1")

setwd("C:/Users/hqv95/Desktop/Data/Wind/Fino1/Long Term/")

horizon <- 1:72
l_horizon <- length(horizon)

# Predictor
usedvar <- "SPEED100"

for (k in 1:l_horizon) { # Loop for every horizon
  RData <- sprintf("fino1_long_term/fino1_long_term_%s.RData", horizon[k])
  load(RData)
  dat <- newdat
  names(dat)[names(dat) == usedvar] <- "X1_t_h"

  # Separation date between training and prediction set
  idx_train <- grep("2015-01-01 00:00:00", dat[,1])
  train <- dat[1:idx_train,]
  eval <- dat[-(1:idx_train),]

  results <- eval[,c("valid_time_t_h","lead_time_h","Y_t_h")]
  results["Y_t_h_pred"] <- NA

  # Fit model with predictor X1 and historical data Yt, Yt-1, ...
  selecttrain <- train[, c("valid_time_t_h","Y_t_h","Y_t","Y_t_m_1","Y_t_m_2","Y_t_m_3","Y_t_m_4", "X1_t_h")]
  GAM_model <- gam(Y_t_h ~ s(X1_t_h, bs="cr") + s(Y_t, bs="cr") + s(Y_t_m_1, bs="cr")
    + s(Y_t_m_2, bs="cr") + s(Y_t_m_3, bs="cr") + s(Y_t_m_4, bs="cr"), data = selecttrain)

  # Predict
  selecteval <- eval[, c("valid_time_t_h","Y_t","Y_t_m_1","Y_t_m_2","Y_t_m_3","Y_t_m_4","X1_t_h")]
  Pred <- predict(GAM_model, newdata = selecteval)

  nam <- rownames(Pred)
l_nam <- length(nam)

for (i in 1:l_nam) {
    results[which(rownames(results)==nam[i]),"Y_t_h_pred"] <- Pred[i]
}

table <- sprintf("GAM_%s/fino1_model_GAM_%s_%s.txt", usedvar, usedvar, horizon[k])
write.table(results, file = table, sep = ",", row.names = FALSE, quote = 1)

RData <- sprintf("GAM_%s/fino1_model_GAM_%s_%s.RData", usedvar, usedvar, horizon[k])
save("results", file = RData)
}
# Compute MAE and RMSE

```r
define working directory
setwd("C:/Users/hqv95/Desktop/Data/Wind/Fino1/Long Term")

# Define horizon
horizon <- 1:72
l_horizon <- length(horizon)

# Initialize empirical vector
empV <- as.vector(matrix(data = 0, nrow = l_horizon, ncol = 1))

# Initialize MAE and RMSE vectors
MAE <- empV
RMSE <- empV

# Loop over horizon
for (k in 1:l_horizon) {
  # Load model RData
  RData <- sprintf("GAM_SPEED100/fino1_model_GAM_SPEED100_%s.RData", horizon[k])
  load(RData)

  # Calculate MAE
  MAE[k] <- mean(abs(results[,3]-results[,4]), na.rm = TRUE)

  # Calculate RMSE
  RMSE[k] <- sqrt(mean((results[,3]-results[,4])^2, na.rm = TRUE))
}

# Print MAE and RMSE
MAE_name <- "MAE GAM SPEED100:
print(paste(MAE_name, paste(round(MAE, digits = 3), collapse=" "))

RMSE_name <- "RMSE GAM SPEED100:
print(paste(RMSE_name, paste(round(RMSE, digits = 3), collapse=" "))

# Save MAE and RMSE
MAE_SPEED100 <- MAE
save("MAE_SPEED100", file = "MAE_GAM_SPEED100.RData")

RMSE_SPEED100 <- RMSE
save("RMSE_SPEED100", file = "RMSE_GAM_SPEED100.RData")
```