# MASTER THESIS

# ASSESSING THE VALIDITY OF MERRA REANALYSIS DATA FOR SIMULATION OF WIND POWER PRODUCTION

Submitted by

**SEBASTIAN MOSSHAMMER**

**0626201**

In partial fulfilment of the requirements for the degree of

**"DIPLOM-INGENIEUR"**

University of Natural Resources and Life Sciences, Vienna

Department of Economics and Social Sciences

Institute for Sustainable Economic Development

Supervised by: Schmid, Erwin, Univ.-Prof. Dipl. Ing. Dr.

Co-Advisor: Schmidt, Johannes, Dipl. Ing. Dr.

Vienna, July 2016

# *Abstract*

The increasing share of electricity generation from wind power is accompanying by installations of wind farms at new and suitable locations. Analyzing the local wind conditions often requires time- and cost-intensive measurements. Therefore, methods are developed to pre-assess the quality of new wind farm locations.

The aim of this thesis is to use MERRA reanalysis data in a wind power simulation model in order to assess the validity of the MERRA data by comparison with real production data. The wind power production of 4 wind farms (1 in Austria and 3 in New Zealand) are compared with the simulation results. At each wind farm, the wind speeds (of the point in the MERRA dataset closest to the wind farm) are extrapolated to the hub height of the turbines by using the empirical derived “power law”. The simulation of the production uses the power curves of the installed turbines (of each wind farm), which represent the relation between wind speed and electricity generation. Several temporal resolutions – from hourly to annual – are analyzed and compared between the 4 wind farms.

Results show that the simulation model does over- and underestimate the total production depending on the location. Correlation coefficients for hourly production are between 0.67 and 0.75 and increase between 0.74 and 0.85 for daily production. In general, low production events are overestimated and high production events are underestimated by the simulation model, although there are exceptions for a production close to rated or zero power. This is a consequence of the low spatial resolution and the utilized smoothed elevation model in MERRA. This could be overcome by the use of an empirical derived optimization model. Hence, the MERRA data has limited suitability for the simulation of single wind farm locations, but it could be useful for the simulation of larger spatial extents.

# *Kurzfassung*

Der steigende Anteil von Windkraft an der Elektrizitätsproduktion bedingt die Errichtung von zusätzlichen Windparks an neuen und geeigneten Standorten. Um die lokalen Windverhältnisse festzustellen, sind meist zeit- und kostenintensive Messungen notwendig. Deshalb werden Methoden entwickelt, um die Qualität von neuen Standorten bereits im Vorfeld abschätzen.

Das Ziel der Masterarbeit ist es MERRA Reanalyse Daten in einem Windkraft Simulationsmodell zu verwenden, um die Validität der MERRA Daten anhand eines Vergleiches mit realen Produktionsdaten zu bestimmen. Die Windkraftproduktion von 4 Windparks (1 in Österreich und 3 in Neuseeland) wurde mit der simulierten Produktion verglichen. Für jeden Windpark wurden die Windgeschwindigkeiten (des am nächsten gelegenen Punktes des MERRA Datensatz zum jeweiligen Windpark) mit Hilfe des empirisch abgeleiteten „Power Law“ auf die Hubhöhe der Turbinen extrapoliert. Im Simulationsmodell wurden die Leistungskurven der installierten Turbinen jedes einzelnen Windparks verwendet, welche den Zusammenhang zwischen Windgeschwindigkeit und Elektrizitätsproduktion darstellen. Verschiedene zeitliche Auflösungen – von stündlich bis jährlich – wurden analysiert und mit allen 4 Windparks verglichen.

Die Ergebnisse zeigen, dass das Simulationsmodell abhängig vom Standort die Gesamtproduktion sowohl unterschätzt als auch überschätzt. Die Korrelationskoeffizienten für die stündliche Produktion liegen zwischen 0.67 und 0.75 und steigen bei einer täglichen Auflösung zwischen 0.74 und 0.85 an. Im Simulationsmodell werden geringere Produktionen überschätzt und höhere unterschätzt, obwohl es nahe der Nennleistung als auch bei null Leistung Ausnahmen gibt. Das ist eine Konsequenz der geringen räumlichen Auflösung und dem eingesetzten geglätteten Höhenmodell von MERRA. Dieses Problem könnte mit einem empirisch abgeleiteten Optimierungsmodell gelöst werden. Die MERRA Daten sind nur bedingt geeignet einzelne lokale Windparks zu simulieren, jedoch dürfte sich diese für größere räumliche Einheiten verbessern.

# *Acknowledgement*

***I would like to thank my family, especially my parents and grandparents, who always strengthened me in my decisions as well as financially and mentally during my whole study.***

***My gratitude also goes to my advisor Erwin Schmid for the provided time and resources.***

***Especially I want to thank my Co-Advisor Johannes Schmidt for the fabulous and excellent support. The calm and highly skilled manner was really impressing.***

***Special thanks to Johannn Baumgartner, with whom I could exchange a lot of ideas since he was working on the same topic.***

***I really would like to thank all my friends and all colleagues at the university for all the good times I was allowed to spend with them.***

***And my very special thank to Lisa Vogler, who was always there for me if I needed help in any way.***

*Table of Contents*

## *List of Figures*

## List of Tables

## R Program-Codes

# 1. INTRODUCTION

As long as we do not discover any new kind of technology that can deliver huge amounts of useful energy, preferably with little or no environmental impact, we will be dependent on the technologies that are known and used nowadays. The rapid development from an agricultural to an industrial society went along with increases in wealth, population, energy consumption, environmental degradation and technological change. Sustainably managing the trade-off between those variables is an important challenge, in particular as wealth is to some extent associated with energy consumption or its availability. Therefore, access to energy should be made universal in a long-term perspective.

Due to the fact that fossil energy sources or even Uranium are finite, technologies that use renewable resources are therefore an important option for the future. The most obvious advantage of renewable energy sources is their renewability, which means "exploiting flows, rather than static resources" [1]. Another advantage of renewable energy technologies over fossil source based technologies is their lower environmental impact, in particular with respect to greenhouse gas emissions, which are strongly related to climate change [2].

Although climate has never been constant in any way – measured in earth´s history – the unusual part of currently observed global warming is that most likely humans are responsible for it. There are several natural effects, like changes in solar activity or the geometry of the earth´s orbit, volcanism, et cetera, that are considered to be the main causes of climate changes [2]. Nevertheless, research results point to a mankind-made global warming due to greenhouse gas (GHG) emissions [1]. Although there are still uncertainties about the accuracy of projections, risks associated with climate change impacts, in particular in the fat tail of impact distributions, are too high. Likewise, geoengineering approaches such as spraying soot in the Arctic or injecting radiation-absorbing dust in the atmosphere are associated with high risks [1]. A less risky way of dealing with climate change is the development of renewable energies.

Since the energy sector is responsible for a huge share of the total GHG emissions, there is a big potential for restricting them. Additionally, it is a sector with a constant high growth rate. A global growth rate of 36% between 2000 and 2010 and a share of about 30% of total GHG emissions are alarming signals especially if a closer look on the energy sector is being taken [3]. Electricity and heat generation is the fastest growing share within the energy sector – from 58.9% in 1970 to 72.6% in 2010. Responsible for this huge share of GHG emissions is the fact that 40.6% of generated electricity comes from coal, 22.3% from natural gas and

4.5% from oil products [4]. This means that more than two thirds of the global electricity generation comes from fossil energy carriers.

For the European Union (EU-28), the share of conventional fossil-based resources in total electricity production in 2015 was 48% and nuclear power had a share of 26%. Quite the same amount as from nuclear power came from renewables, whereby 12% of the total electricity production came from hydro, 10% from wind and the last 4% are spread amongst several renewable technologies like photovoltaic, biomass or geothermal. One of the highest growth rates within the renewables comes from wind power. From 2014 to 2015, an increase of 21.8% within the EU-28 shows the importance of wind power as a technology to replace fossil energy based technologies [5]. It is quite obvious that wind power is one of the most promising paths to generate electricity in a sustainable way at relatively low costs.

Building new wind power plants or wind farms requires, amongst many other factors, locations that can provide good wind conditions. Beside the country specific feed-in tariffs and subsidies, the wind conditions are probably the most important factor for the successful operation of a wind farm. Since measuring wind speeds at different heights and locations is costly and time-intensive, an alternative, time and cost saving method to figure out the potential of electricity generation would be very useful. The method presented in this thesis uses long term wind speed data from reanalysis datasets to develop a simulation model and compare afterwards the simulated electricity generation with the real electricity output. Four wind farms, 1 in Austria and 3 in New Zealand are examined. The wind speed data is taken from the MERRA-project, which is a reanalysis product developed by NASA.

The main goal of this thesis is to examine how well the MERRA-data perform on a local scale with regard to simulating hourly wind power production. The MERRA-Dataset (Modern Era-Retrospective Analysis for Research and Applications) is a reanalysis product, which processes meteorological observations, respectively records from satellites or conventional observations (e.g. dropsondes, radiosondes, PIBAL winds, wind profiles) in the past, and interpolates them on a global grid with 1/2° latitude, 2/3° longitude and 72 vertical levels. In sum the MERRA-grid consists of 183.600 points - 540 points longitudinal times 340 points latitudinal [6]. It was developed by the Global Modelling and Assimilation Office (GMAO) from NASA. A primary objective is to put observations from EOS (NASA´s Earth Observing System) satellites into a climatic context. The observations and the reanalysis reach back to 1979, whereby MERRA was developed in 3 different stages, respectively data-streams. The current stream runs since 2001 and delivers a nearly real-time tool for climate analysis. Several different products are available and can be accessed online [7]. The product and files used here are described more precisely in chapter 2.2.1./2.2.3.

The advantage of using reanalysis-data compared to measured data is on the one hand that the data is available without any measurements gaps since 1979 and on the other hand that the data is available globally and if interpolated horizontally for each desired location. The available time-span is growing every day, as the data is continuously updated. Also, unlike real production data, wind speeds can be used to simulate the production for different types of turbines. If data quality is sufficient, a simulation model can be developed to reproduce the wind speeds or respectively wind power generation for any location with a much lower effort than processing measured data. Those time series can be used to detect “wind-hot-spots” or obtain a pre-assessment of a potential wind farm location. Also, they can be used in large scale integration studies to assess the optimal integration of wind power in the power system.

The simulation of wind power production by means of wind speeds of the MERRA-data and the specific power curves of the turbines used in the wind farms should show comparable statistical characteristics to the real production. Hitherto, most of the studies which used MERRA data for simulating wind power or photovoltaic production or comparing wind speeds focused on larger areas, e.g. [8]–[10]. In order to validate the data and to estimate the value of the MERRA-data for this specific application it has to be figured out where the weaknesses or strengths are. Are productions or respectively wind speeds over- or underestimated? How good can the simulation for different temporal resolutions reflect the real production? Are there differences with regard to the time profile or wind speeds, respectively electricity generation? Are seasonality and the time profile comparable? In order to answer these questions, a comparison of the simulated production with real production data of 4 wind farms is carried out.

Of course there are known open issues with respect to the reanalysis data, e.g. they are time averaged, the topology is smoothed and horizontal interpolation is applied to the MERRA data [7]. These issues have, of course, an influence on the quality of the data and furthermore the results. Nevertheless the method used and presented here should provide an alternative to measuring wind speeds conventionally. Future extensions can be implemented later to further increase the quality of simulation results, as further improvements are not in the focus of this thesis.

Chapter 2 takes a look at the characteristics of renewables and their role in energy system modelling, especially for systems with a high share of renewables. This chapter closes with a brief review of studies that used MERRA-data. Chapter 3 describes the used data and the methods to manipulate the data as well as the development of a simulation model that uses MERRA-data to simulate hourly production of 4 wind farms. It closes with the methods used for analysing and evaluating the results of the simulation model. Chapter 4 shows results and the evaluation of the simulation model. Beside the analysis of the hourly data, several time spans are aggregated and analysed with regard to statistical significance. The thesis closes with a discussion and the conclusions of the results.

# 2. REVIEW ON THE USE OF REANALYSIS DATA

Several studies have already used reanalysis data to model the integration of renewables in energy systems. Chapter 2.1 shortly describes the basics of energy system models and discusses a few studies that used reanalysis data for developing a model which reflects an ideal mix of renewable energy sources for different areas, respectively countries. Chapter 2.2 discusses a few studies that used MERRA data for simulating electricity generation by wind power or photovoltaics, whereby these studies also focus on whole countries. In contrast, the analysis of the quality of MERRA data for simulating single wind farms is assessed in this thesis.

## 2.1. MODELLING ENERGY SYSTEMS WITH LARGE SHARES OF RENEWABLE ENERGIES

Energy conversion- or electricity generation from renewable resources, especially from wind, direct sun radiation or water, is dependent on climatic conditions. The annual hours of sunlight, the amount of precipitation or wind conditions determine how much electricity from renewable energy sources can be produced and at which moment. A huge challenge for energy systems that are mainly based on renewables is the volatility of production due to the dependency on natural processes. This can be a serious problem for the security of supply. Where, how and at which moment electricity can be produced or stored in the most efficient way, will be a huge and increasing challenge in the future.

To deal with it, highly accurate and solid data about solar irradiance, wind or precipitation on site make it possible to create a regionally adapted infrastructure, which can generate electricity as stable and predictable as possible.

Energy system models are principally based on several data inputs, which represent the present state of the system or project system states into the future. They include factors like pricing, characteristics and availability of technologies or energy carriers or fluctuations of generation in time or the total demand and supply of energy respectively electricity. The fluctuations of generation in time from renewables due to the dependency on meteorological conditions have a serious impact on energy system models, because they affect the optimal expansion of power plants, storage infrastructure or grids [11]. Hence it is very important to estimate the generation over different time periods as well as possible. Several studies are available (for example [12], [13]) that assessed optimal mixes of renewables in the electricity supply. Most energy system models are made for an hourly resolution. For modelling the generation of renewables, different input data is used, like real feed-in data from generation, measurements from weather stations, satellite measurements or reanalysis data [11].

A study [14] from the Harvard China Project claims that the whole electricity production of China could be covered by wind power for only "slightly higher costs" than from fossil fuels but they only took annual production into account and did not match demand and supply continuously. Backup power, storage or transmission infrastructure was not considered, which is necessary to balance fluctuations. Hence, Huber et al. [13] developed a model to figure out the optimal mix of variable renewable energy sources (VRE) for China in an hourly resolution. MERRA data was used to calculate the hourly wind and PV generation for all regions of China. After calculating the hourly electricity demand, several scenarios with varying shares of VRE were developed, resulting in different residual loads, capacity credits and storage requirements. It was shown that 20% of the total electricity consumption can be supplied by VRE, resulting in nearly no positive residual loads at all and hence no storage facilities needed to be installed. The problem arises with higher shares of VRE: If wind and PV would produce 100% of the annual consumption, only 50-75% of the demand on hourly basis could be covered due to the lack of storage possibilities. The optimal mix of wind and PV depends on the installed capacity of VRE. For shares of VRE over 50% "defining the optimal mix becomes complex and uncertain" since the storage requirements increase to non-viable high values [13]. Nonetheless they come up with a suggestion: Planners should pursue a mix of 70% wind and 30% PV.

Another attempt was made by Schmidt et al. [12], in which the authors assessed the optimal mix of PV, wind and hydro power for Brazil in 2034 in order to achieve a low carbon electricity supply and avoid an increase of thermal power generation. It is shown, that a mix of wind, PV and hydro power would reduce risks compared to a hydro-thermal only system. An optimization model determines the optimal mix of wind power, PV and hydro power by simultaneously reducing the thermal power dispatch. Afterwards a simulation model was used to assess the reliability of the system. In other words, the authors assessed if demand and supply can be matched in an operational model without foresight by using the power mix determined by the optimization model. The data used for the simulation of PV, wind power and hydro power production includes solar irradiation, wind speeds and water inflows. Wind speed data for example was used from the ECMWF (European Centre for Medium-Range Weather Forecasts) with a spatial resolution of 0.75x0.75 degrees, 3 hourly intervals from 1979 to 2014 and from NCAR/NCEP (National Centres for Environmental Prediction and the National Centre for Atmospheric Research) with a spatial resolution of 2.5x2.5 degrees, 6 hourly intervals and a time period from 1948-2014. Validation showed that the data from ECMWF could reproduce solar irradiation and wind speeds best and was therefore used for the whole model.

In this thesis, reanalysis data from NASA (MERRA-reanalysis) is used and validated. The MERRA dataset was used because the necessary parameters are available in hourly-resolution. Another product would have been the ERA-Interim from the ECMWF (European Centre for Medium-Range Weather Forecasts) but the time resolution is lower at 3-hours .

With the help of reanalysis, all kind of meteorological parameters can be reconstructed, which have not been measured, whereby information or measurement gaps can be closed.

Hence this offers an instrument to develop models for energy production systems or at least reanalysis data can be integrated in an energy production model. Therefore it is crucial to know how well the reanalysis data perform and how reliable they are.

## 2.2. STUDIES ASSESSING MERRA DATA QUALITY

There have already been made a few studies which used reanalysis- or especially MERRA-data to - simulate wind velocities, wind power or photovoltaic production and compare the simulation to real production data. Some examples can be found in references [8]–[10], [15], [16].

A very well performing model was developed by Bergkvist and Olauson [8] which compared data of total Swedish power generation, taken from the Swedish TSO (Transmission System Operators) with simulated data, using MERRA-data. A mean absolute error in hourly energy of 2.9% and a RMS (note: root mean square) error of 3.8% showed the good performance of the data set. The authors used power curve smoothing and bias correction to achieve these results. The power curve smoothing included “higher power around cut-in wind speed, lower power around rated wind speed and a more smooth transition from rated to zero power at cut-out wind speed”. The power curve was made as a function of the incoming wind (power) and therefore external and internal losses could be figured out much better. The bias correction included seasonal and diurnal bias for the aggregated production. It is worth to mention that the low errors are only valid for the whole of Sweden and not for sub-areas or single wind farms. The simulation of smaller areas resulted in larger errors.

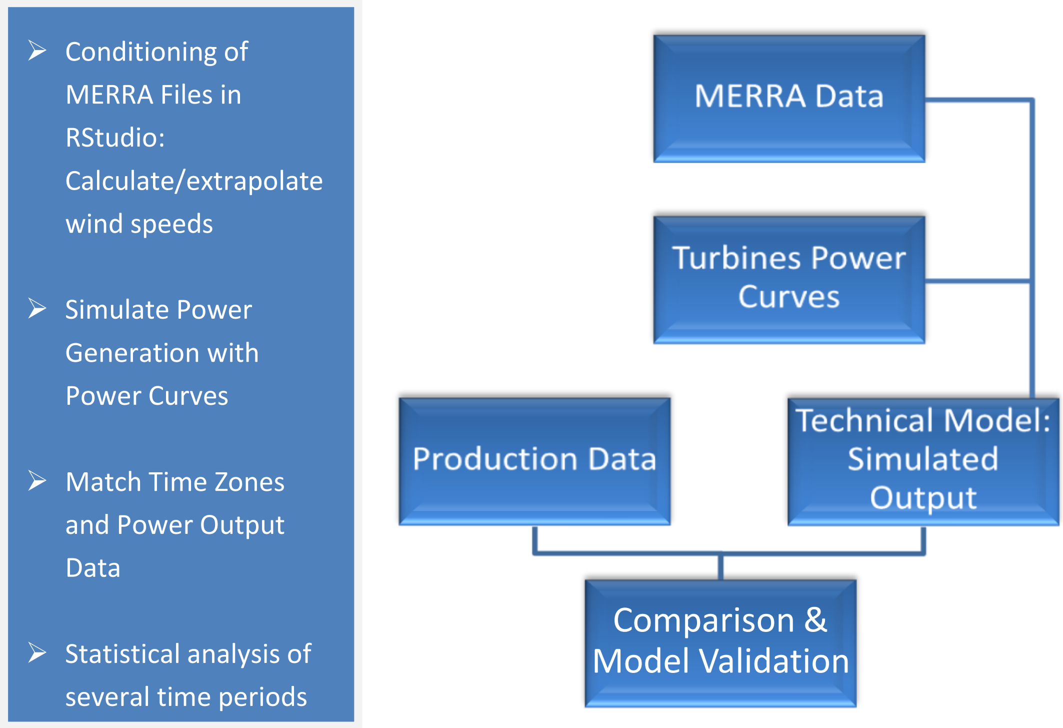
Ritter et al. [16] compared wind speeds from 7 wind farms with wind speeds from MERRA-data and developed a production map that illustrates the estimated yearly production potential for whole Germany. This could be useful for governments, operators or investors as a pre-assessment and to check the suitability of a location.

A publication of Cannon et al. [9] demonstrates that the MERRA reanalysis data cannot only be used for long-term investigation but also to estimate the frequency of (short-lived) extreme events. Wind speeds of MERRA and MIDAS (Met Office Integrated Data Archive System – containing meteorological data from the United Kingdom) were compared and the results showed that MERRA data generally underestimates high wind speeds and overestimates low wind speeds. The underestimation for high wind speeds is slightly removed if only stations up to 300m altitude are taken into account, which is a consequence of smoothing topography in MERRA-data. However, mean wind speeds are reproduced quite accurately. Extreme events, which are classified by the CF (capacity factor) into 3 thresholds for high and low CF´s, achieve a quite good agreement between MERRA and MIDAS. However, long-lasting events are usually overestimated whereas short-lasting events are mostly underestimated by MERRA. Even the rarest and most extreme persistent events could be reproduced quite well in most of the cases. Nevertheless this applies only to aggregated-data for the whole of Great Britain. And it is also mentioned that for ramps in production, MERRA tends to underestimate these ramps, therefore the authors suggest applying dynamical downscaling to the MERRA data.

MERRA data can also be used to model photovoltaic power production as Juruž et al. [10] demonstrate. In this study, shortwave fluxes from different data-sources – in-situ measurements, HelioClim-1 and HelioClim-3 – were compared with MERRA-data. Although MERRA overestimated irradiance, the correlation is improved after a bias correction. The authors conclude that MERRA is useful for studying interannual variability for medium and long-term planning of photovoltaic production.

# 3. DATA & METHODS

The primary objective of this thesis is to determine how well the data of the MERRA reanalysis performs in the simulation of wind power production. The method which was used here is illustrated step by step in figure 3.



➢ Conditioning of MERRA Files in RStudio: Calculate/extrapolate wind speeds
➢ Simulate Power Generation with Power Curves
➢ Match Time Zones and Power Output Data
➢ Statistical analysis of several time periods
MERRA Data
Turbines Power Curves
Production Data
Technical Model: Simulated Output
Comparison & Model Validation

Figure 1: Graphical Overview of the Validation Process, Source: Own Figure

The MERRA data is prepared in R-Studio in order to calculate wind speeds, interpolate them to the location of the wind farm and extrapolate them to the hub height of the turbines. Afterwards the electricity production can be simulated for each wind farm by using the power curves of the installed turbines. The result is a simulated hourly production for 3 Wind farms in New Zealand and 1 in Austria. A comparison and statistical analysis of the technical model with the real production data should show how well the MERRA-data performs. The statistical methods for comparing real and simulated production are described in chapter 3.3 and the results are presented in chapter 4. The simulation model was developed with R-Studio Version 0.99.491 plus the additional packages “lubridate”, ”ncdf4”, ”plotly” and “psychometric”. The R-version, necessary for R-Studio to work, was 3.2.5.

## 3.1. Real Production Data

The examined wind farms are shortly described in the subsequent chapters. The characteristics of the installed turbines are shown and the production data is examined to remove data omissions or errors. Production gaps and erroneous data could bias the results and were dropped therefore.

There is one production data-file for each region – 1 file for Austria with 1 wind farm and 1 file for New Zealand with 7 wind farms, whereby only 3 are examined. The file for New Zealand contains the hourly production for those 7 wind farms with different time spans because the wind farms have been built at different times.

### 3.1.1. Wind farm in Austria

Due to data protection requirements, there cannot be given precise information about the wind farm in Austria. The operator, exact production information, number of turbines as well as the exact location cannot be made public.

The existing data include the records of a power feed counter in which the electricity production of the wind power plants, using Enercon E70-E4 turbines with a hub height of 86m and 2.050kW maximum power each, are recorded. The records reach from 01.10.2006 at 0:00 to 17.12.2012 at 24:00 and represent the quarter-hourly production. In the first 10.000 hours, the mean in production is significantly lower than in the remaining hours, as shown in Figure 2. The reason is not known, so the data was dropped to not distort results. [1]

Due to the fact that the simulation is done in hourly intervals, the real production data is aggregated to full hours, e.g. from 00:00 to 01:00, 01:00 to 02:00.

The MERRA-data are collections which consist of a sequence of data averaged over an interval and a certain time, which in this case is an hourly interval at 00:30 GMT, 01:30 GMT, 02:30 GMT, therefore it makes sense to aggregate the real data to full hours, so that the MERRA-data behaves analogous to the real data.

[1] The production data in Figure 2 were normalized over the maximum production due to data protection requirements.

[2] It has to be mentioned that averaged values cannot reflect the production exactly due to the non-linear

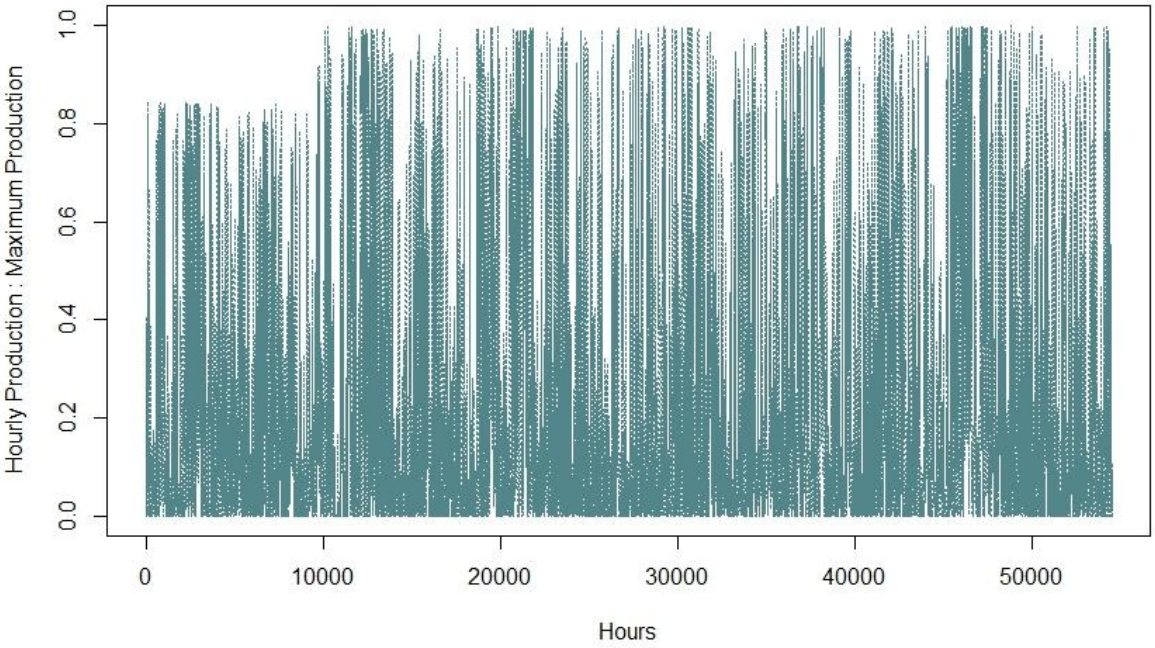


Figure 2: Normalized electricity production of Austrian wind farm from 01.10.2006 to 17.12.2012, source: own diagram

### 3.1.2. WIND FARMS IN NEW ZEALAND

3 Wind farms have been studied: White Hill, Mahinerangi and Te Apiti. For the wind farms in New Zealand, there are no data protection restrictions.

The dataset for New Zealand includes the hourly production for each wind farm in MWh. The time period for the production data for each wind farm is different, because they have been built or put into operation at different times. The time series of the wind farms in the used Excel-sheet all end at the same date, 31.03.2013.

The choice for the 3 wind farms (out of the 7 existing datasets for all wind farms in New Zealand) has its reasons in the geographical location of the wind farms and the availability of further information about the wind farm, for example installed turbines or information about the turbines. No power curve could be found for the installed turbines of the Tararua wind farm and it is located very close to Te Apiti, so it was not considered for further studies. For Te Rere Hau it was not possible to figure out, which turbines are installed. Project West Wind and the Te Uku wind farms were also an option but the decision fell on Te Apiti because of the long production data (the longest of all wind farms) and its location on the North Island. White Hill was put into operation on June 2007 so there was also a long production data set available and it is unlike Te Apiti located on the South Island. Mahinerangi has a rather short period of production data, but it was chosen as the third wind farm because it´s almost on the same latitude like White Hill, which could results in a better comparison of these 2 wind farms. Figure 2 shows the location of the 3 wind farms in New Zealand.

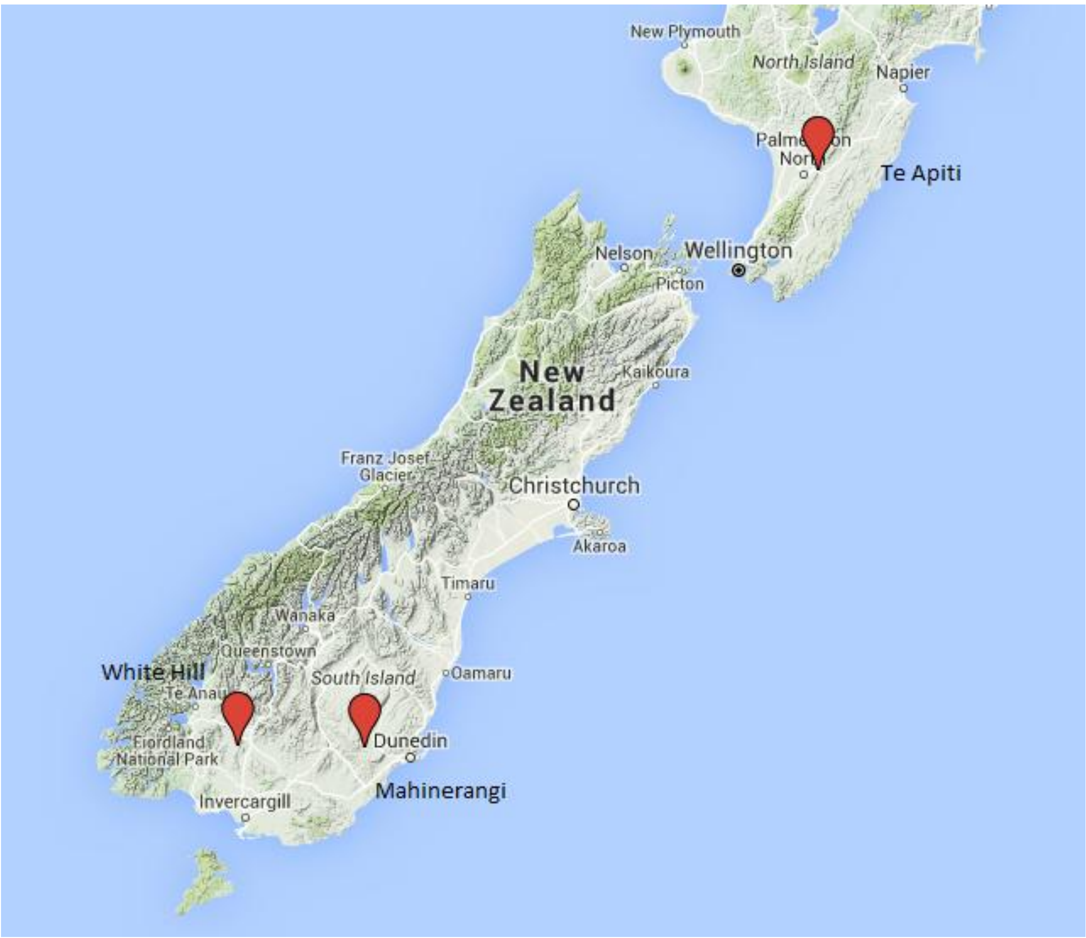


Figure 3: Location of the studied wind farms in New Zealand, source: maps.google.com

### 3.1.2.1. WHITE HILL

This wind farm consists of 29 wind power plants using Vestas V80-2.0MW turbines with a hub height of 68m. The operator is "Meridian Energy". The point used for the simulation has the coordinates $-45.7525°\mathrm{N}$ and $168.271667°\mathrm{E}$ [17].

All together the wind farm has a capacity of 58MW. The park started to operate in June 2007. Figure 3 shows that it took about 3.000 hours to reach full production. A lack of wind is not very likely, as the increase in the mean is very regular and the MERRA wind speeds are consistently above 10m/s up to more than 15m/s. Compared to the subsequent production, some sort of a start-up process is the most likely event. So the first 3.000 hours are excluded from the dataset. Also the time span from 31.000h to 32.750h is excluded because it´s assumed that there was some maintenance processes, since in most of this time there has not been any production. A similar event is not observed in the remaining periods and due to the fact that the MERRA wind speeds during this time are consistently above 5m/s and go up to nearly 20m/s, it is very likely that it was not a lack of wind that caused this no-production period. As the operator states, the location suits very well for a wind farm, since the “Southland” area has strong, constant winds [18].

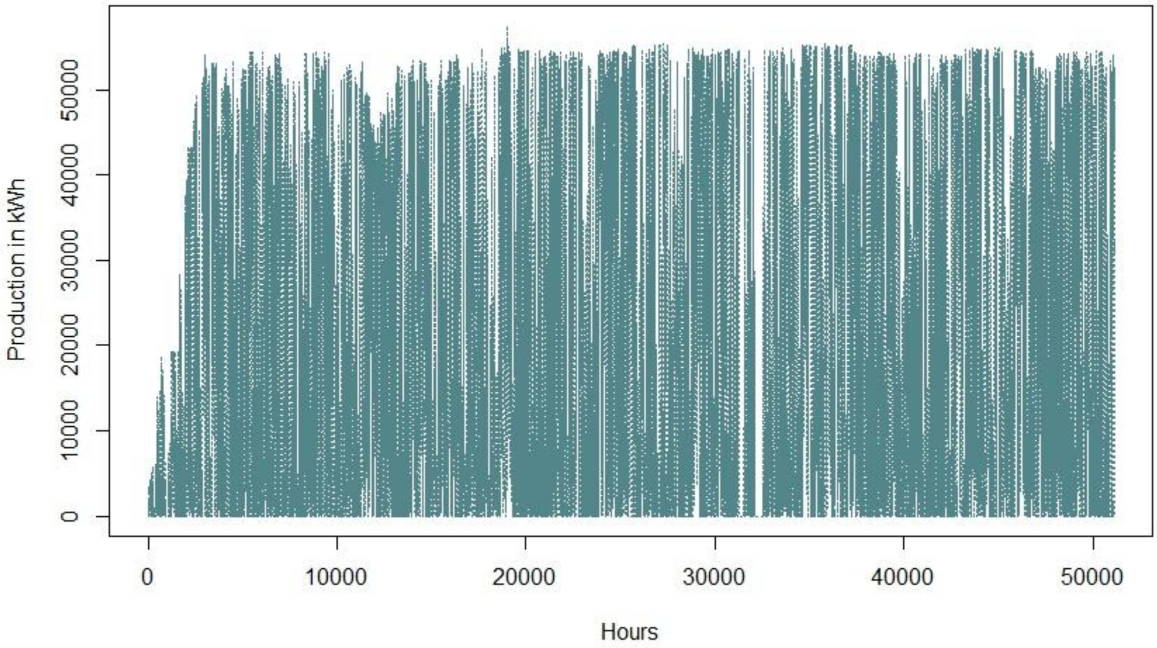


Figure 4: Electricity production White Hill Wind Farm from 01.06.2007 to 31.03.2013, source: own diagram

### 3.1.2.2. Mahinerangi

The operator of the wind farm is “TrustPower”. It is located approximately 70km west of Dunedin and started production in March 2011. It is the wind farm with the shortest production time. It consists of 12 wind power plants with Vestas-V90 3MW turbines each with a hub height of 80m. Until now it has only 36MW of installed capacity, which could be extended to 200MW due to a permission of the operator [19] [20].

The production data has some gaps in it, which can be seen in figure 4, that´s why the first 1.200 hours, the time period between production hour 3.200 and 3.500 and between 4.600 and 4.800 are not examined. The first 1.200 hours are assumed to reflect a starting process since the mean production is significantly lower than in the periods afterwards. The gaps with mostly zero production between hour 3.200 and 3.500 as well as between 4.600 and 4.800 are assumed to exist because of maintenance processes, since the reanalysis data show wind speeds from 3 to 12m/s during this time, which would be more than sufficient wind speeds for a normal production profile.

The whole wind farm extends on an area of 17.23km² - the point used for the simulation has been taken from [21] and has the coordinates $-45.760556°N$ and $169.905°E$.

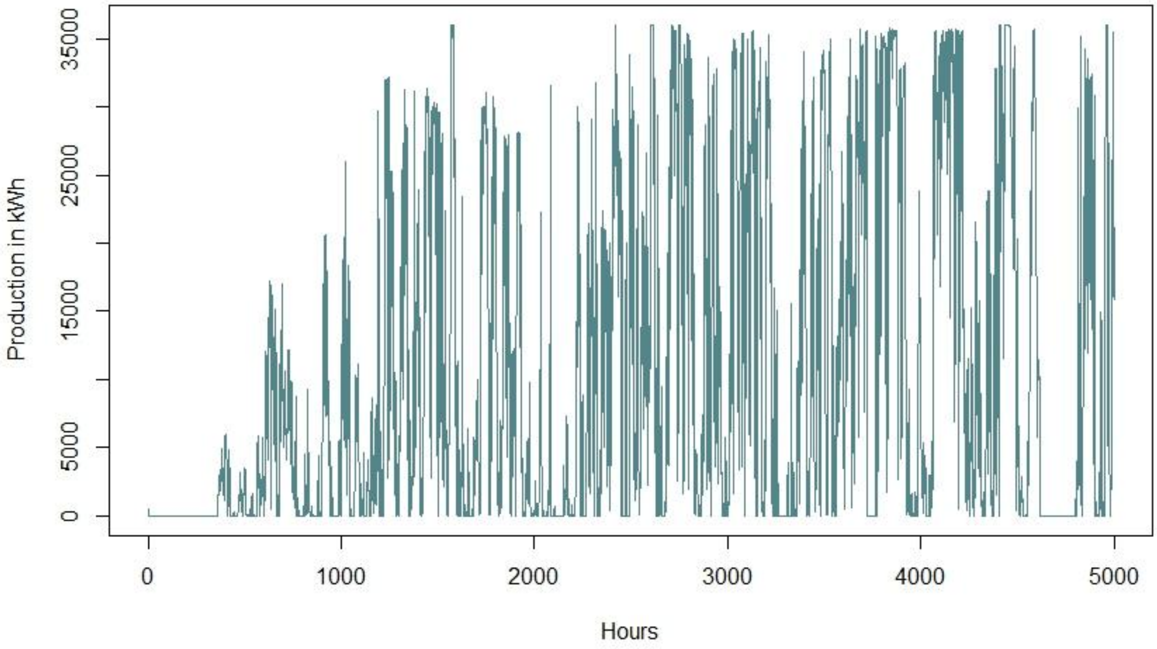


Figure 5: Part of the electricity production in the Mahinerangi Wind Farm from 04.02.2011 to 31.03.2013, source: own diagram

### 3.1.2.3. Te Apiti

Since 26.07.2004, electricity is produced in this wind farm. The simulated period starts on 01.01.2005, therefore the production of 2004 is not been taken into account. It´s still the longest observation period of all examined wind farms. It´s located on the North Island of New Zealand, north-east of Palmerston and north of the “Manawatu” gorge. The extraordinary wind conditions, even for international standards, are because of the “Manawatu” gorge, which functions as a wind funnel, as mentioned in [22]. The wind farm extends on an area of 11.5km².

The point which was used for the simulation has been taken from [23] and has the coordinates $-40.296111°N$ and $175.808333°E$.

This wind farm was the first that fed the electricity not only in locally but also in the national transmission grid.

It consists of 55 wind power plants using Vestas / Micon NM72-1650 turbines, each with a power of 1650kW and a hub height of 70m. The whole wind farm therefore has a capacity of 90.75MW. The turbine was originally constructed by Micon, but due to a fusion of Micon and Vestas, Vestas is now listed as manufacturer [22], [24], [25].

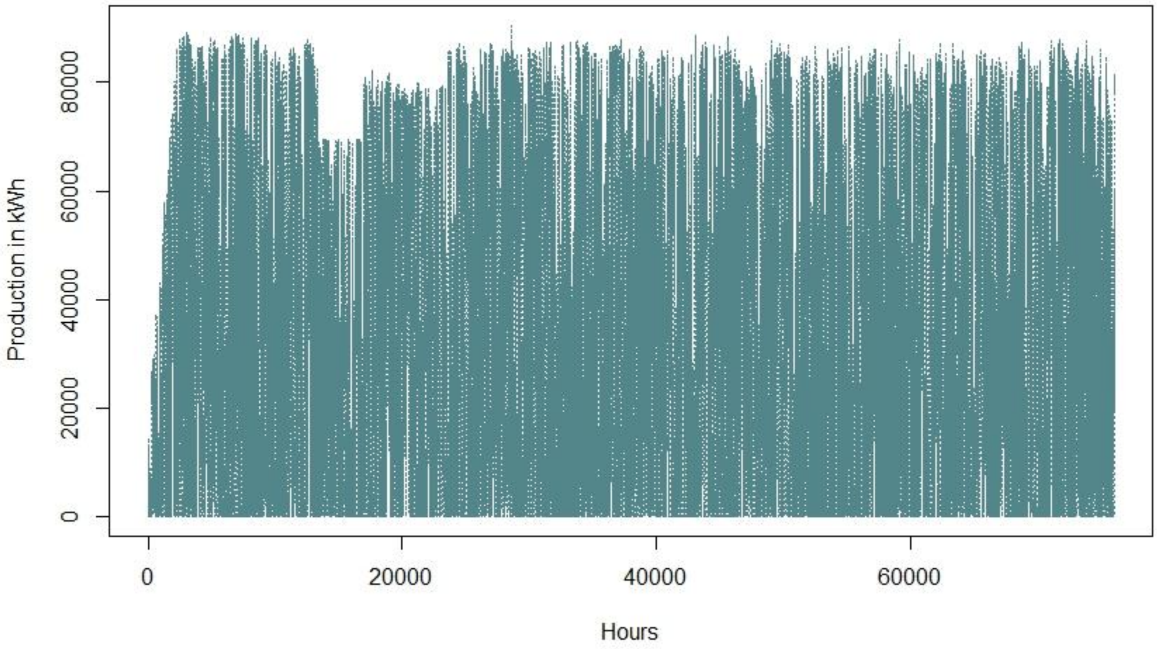


Figure 6: Electricity production of Te Apiti Wind farm from 26.07.2004 to 31.03.2013, source: own diagram

As Figure 5 shows, there is a time period with reduced production, for unknown reasons, from production hour 13.500 to 24.000. This time span is not examined and is removed from the dataset. And, as mentioned before, the production from 2004 is also not used because the MERRA dataset downloaded starts in 2005, hence the starting-up process of the wind farm, which can be observed in Figure 5, is also removed.

## 3.2. A Simulation Model for Wind Farms

This chapter describes the necessary steps to develop a model that can simulate the wind speeds respectively the production of the 4 mentioned wind farms with the MERRA-data. At first, the MERRA-files are described and the relation between wind velocity and electricity production is mentioned. In order to use the MERRA-files, they have to be prepared, which is done in RStudio. Functions for reading the files, as well as the extrapolation of the wind speeds from the given height of the MERRA-files to the hub height of the turbines are described. Afterwards, a data-frame with the extrapolated wind speeds, the real production and the matched timestamps is generated in order to be able to simulate the production by means of the power curves and wind speeds.

### 3.2.1. Access and Characteristic of MERRA-Files

There are several products of MERRA offered and available on http://disc.sci.gsfc.nasa.gov/daac-bin/DataHoldings.pl. First the product was chosen, afterwards the access-method "Data Subsetter" and further "Daily Product".

The access can also be done directly on http://disc.sci.gsfc.nasa.gov/daac-bin/FTPSubset.pl?LOOKUPID_List=MAT1NXSLV. For this study, the product "IAU 2d atmospheric single-layer diagnostics" was used. Further information is given in chapter 2.2.2.

Two Datasets were obtained – one for New Zealand and one for Austria. The spatial boundary for the Austrian dataset is between $46°N$ and $50°N$ latitude and $12.66°E$ and $20°E$ longitude. The New Zealand dataset has the spatial boundaries $-48.5°N$ to $-32.5°N$ latitude and $165.33°E$ to $180°E$ longitude. The wind farms are all located between these boundaries and the boundaries were chosen generously because the used interpolation method was not set before the download of the data.

The time span was determined from 01.01.2005 to 31.12.2014 for each dataset and the chosen parameters were:

```
U10M ........ Eastward wind at 10 m above displacement height
U2M ......... Eastward wind at 2 m above the displacement height
U50M ........ Eastward wind at 50 m above surface
V10M ........ Northward wind at 10 m above the displacement height
V2M ......... Northward wind at 2 m above the displacement height
V50M ........ Northward wind at 50 m above surface
DISPH ....... Displacement height
```

“NetCDF” was set as the output file format. After the file search is finished, a text file with a list of URL´s is available. Each URL represents a file and can be downloaded manually or automatically with the application “wget”. Therefore it is necessary to copy the text file into the folder in which “wget” was installed and afterwards type in the command “wget --content-disposition –i >name of the text file<” into the Windows command prompt and the download process starts. For the chosen time span of 10 years, this results in 3652 files respectively days.

The files contain hourly averaged values for the central time of the hourly interval. This means the times are 00:30, 01:30, 02:30 etc. GMT (Greenwich Mean Time). The time shift to the locations of the wind farms, and the presence or absence of day light saving time, have to be considered. This is shown more precisely in chapter 2.2.3.

One file contains the values of the chosen parameters for each hour of a day and each point of the MERRA-grid that is located between the determined boundaries of the datasets. For the Austrian dataset, this results in 12 points longitudinal and 10 points latitudinal which means there are 120 points. For the New Zealand dataset this results in 759 points, 23 points longitudinal and 33 points latitudinal. This implies that the Austrian dataset covers an area of 8 degrees longitude and 5 degrees latitude, because the MERRA-grid has a resolution of 0.5° latitude and 0.66° longitude. The parameters are wind vectors (U/V) on the one hand and the displacement height on the other. The displacement height is described as an increased surface due to the vegetation on site, the covering of snow or buildings, as those elements in the landscape act like a resistance for the wind. For vegetation the displacement height varies from 0.4 to 0.8 from the average vegetation height. It depends on the density and type of vegetation. For snow the displacement height is the same as the height of the snow coverage [26]. Since the height and density of vegetation changes in the course of the year, the displacement height is a variable parameter, which changes during the year. The variability of this particular parameter is not being validated in this thesis.

### 3.2.2. Wind Velocity and Electricity Production

Like mentioned above, $u$ and $v$ are the wind vectors. The wind vector $u$ describes the wind in eastward direction, the wind vector $v$ the wind in northward direction in $m/s$. The second part of the parameter name is the height – 2m above displacement height, 10m above displacement height and 50m above surface. Hence the parameter "U2M" for example explains itself as "eastward wind 2m above displacement height".

The wind velocity results in applying equation (1) to the wind vectors:

$$v = \sqrt{u^2 + v^2} \qquad (1)$$

After applying equation 1, the wind direction is not traceable anymore, but this is not necessary for the simulation because only the wind velocity is needed. It is assumed that each wind power plant can actively control it´s direction and therefore be able to face the wind directly.

The following equation describes how the power of the wind can be calculated:

$$P_{Wind} = \frac{1}{2} * \rho * A * v^3 \qquad (2)$$

With:

$P_{Wind}$ ........ Power of the wind

ρ ........... Air density

A ........... Wind flown through area

v ........... Wind velocity

Most notablely in equation (2) is that the wind power increases with the third power of the wind velocity. A wind power plant can transform a part of the kinetic energy of the wind into electricity. The power coefficient $C_p$, which depends on the installed turbine, describes how much of the energy contained in the wind can be transformed by the wind power plant into electricity. The theoretically highest usable power coefficient is described by the Betz´ Law and is 59.3% of the total energy contained in the wind. A modern wind power plant can, under perfect conditions and between a certain wind speed interval, harvest about 50% of the total energy contained in the wind, which is quite close to the theoretical maximum [2]. Since the power coefficient is not needed for this simulation model, it´s not discussed any further.

Shadowing effects – the fact that in a wind farm single wind power plants can affect each other by slowing down the air after passing the turbines blades and hence deliver less power – are not being taken into account for the simulation.

The generated power as a function of wind speed can be plotted as a power curve. The data about turbines can mostly be found on the operators' websites or elsewhere on the internet. Figure 3 shows the power curves for all turbines that are used in the examined wind farms.

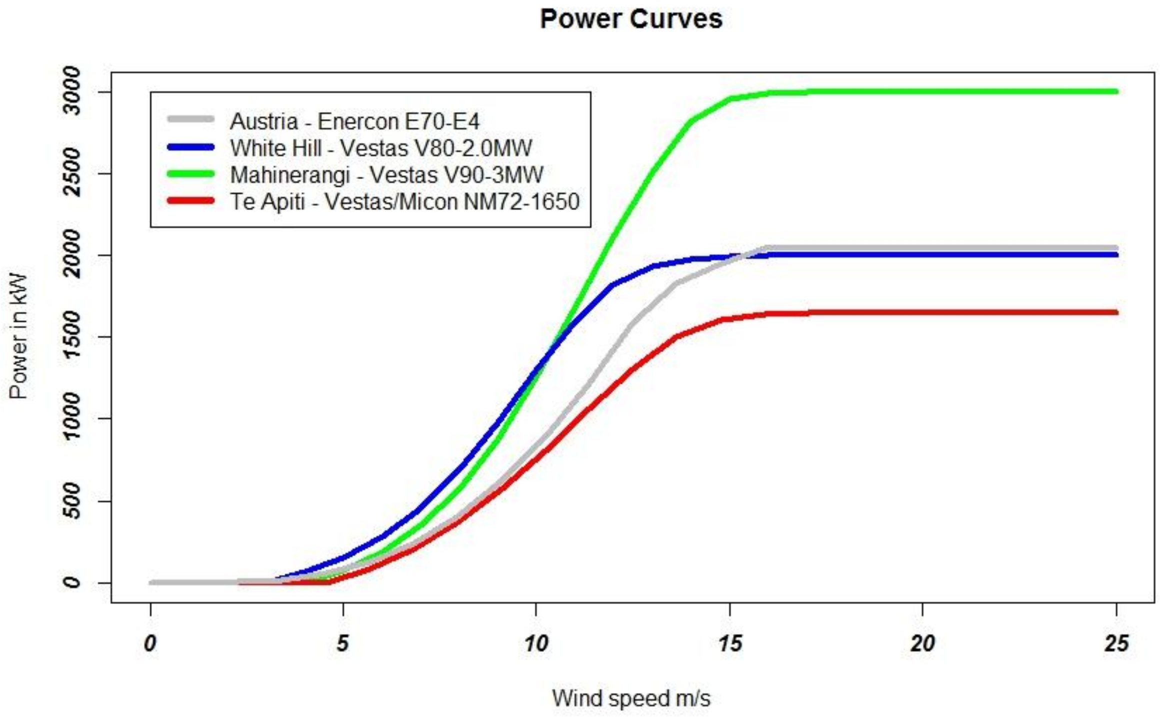


Figure 7: Power curves for 4 different turbines used in the examined wind farms, sources: [27]–[30], own figure.

It can be observed that the turbines have a similar curve, differing mainly in the maximum power output. The Vestas/Micon NM72-1650 and the Enercon E70-E4 have a weaker performance between around 6m/s and 14m/s. Especially the Enercon performs worse compared to the Vestas V80-2.0MW turbine, although they have a similar maximum power output. Usually the turbines start to produce power at around 4m/s, this is called the cut-in wind speed. The produced power then increases from around 5m/s to 12m/s, at which speed they nearly reach maximum capacity. If the wind speed reaches more than 25m/s, the turbines shuts down (cut-out wind speed). For example, the Enercon E-70-E4 turbine has a cut-in wind speed which is indicated at 2.5m/s and cut out wind speed at 28-34m/s [31]. But since there are no wind speeds above 25m/s in the MERRA-datasets, the curves are only displayed up to 25m/s wind speed. The data for the power curves were derived from: Enercon E70-E4 [27], Vestas V80-2.0MW [28], Vestas V90-3MW [30] and Vestas/Micon NM72-1650 in [29].

### 3.2.3. Preparation of MERRA-Data

First, MERRA files had to be downloaded, as explained in section 3.2.1. The filename of a MERRA-file is explained here exemplary:

$$MERRA300.prod.assim.tavg1\_2d\_slv\_Nx.20050101.SUB.nc \tag{2}$$

With:

| | |
|---|---|
| MERRA300 .... | Part of the original third and actual data-stream |
| prod.assim .. | Product of assimilation stream |
| tavg1 ....... | time averaged – data consists of 1-hourly averaged values |
| 2d_slv ...... | data is 2-dimensional and single-level |
| 20050101 .... | date 01.01.2005 |
| .nc ......... | data file output format NetCdf – Network Common Data Format |

More detailed explanations to these and other abbreviations can be found in [6].

An important observation has to be made with respect to some of the downloaded files: some of the MERRA-data was reprocessed for a limited time period, as a compiler was improved and, at the same time, program code was updated. Hence, the data from 01.06.2010 to 31.07.2010 was reprocessed and marked with “MERRA301” instead of “MERRA300”. This is described in detail in [32].

The data used here contains the mentioned time span, so it was necessary to rename the files from “MERRA301” to “MERRA300”. Renaming was done with the application “Rename-Master”. This was necessary because otherwise the data wouldn´t be in the correct chronological order when used in RStudio.

#### 3.2.3.1. Reading Functions in RStudio

To be able to use the data in RStudio, an additional package is needed. During this thesis, the package “ncdf4” was used. With the developed reading functions for NetCdf-files, the

MERRA-files can be opened and then the used parameters are extracted. With the following reading function the parameter “u50m”, eastward wind in 50m above surface, is extracted:

```
readu50m <- function(ncname) {
  du50m <- "u50m"
  ncfile <- nc_open(ncname)

  #Longitude
  longitude <- ncvar_get(ncfile, "longitude", verbose = F)
  nlon <- dim(longitude)

  #Latitude
  latitude <- ncvar_get(ncfile, "latitude", verbose = F)
  nlat <- dim(latitude)

  #Time
  time <- ncvar_get(ncfile, "time")
  tunits <- ncatt_get(ncfile, "time", "units")
  ntime <- dim(time)

  #read the variable
  u50m.array <- ncvar_get(ncfile, du50m)

  #Dataframe
  u50m.vec.long <- as.vector(u50m.array)
  u50m.mat <- matrix(u50m.vec.long, nrow = nlon * nlat, ncol = ntime)
  lonlat <- expand.grid(longitude,latitude)
  lonlat <- lonlat*rad

  # Distance between points
  dista <- 6378.388*acos(sin(lat) * sin(lonlat[,2]) + cos(lat) *
cos(lonlat[,2]) * cos(lonlat[,1]-long))
  dfu50m <- (data.frame(cbind(dista,lonlat/rad, u50m.mat)))
  dfu50m <- dfu50m[ order(dfu50m[,1]), ]
  names(dfu50m) <- c("dista", "Longitude", "Latitude", seq(1:24))
  nc_close(ncfile)
  return(dfu50m)
}
```

R Program-Code 1: Reading function for “U50m” parameter, source: own Code

The developed function is called “**readu50m**” and is saved as such. The curly brackets mark the beginning of the function per se – what it contains and what it should do later with one or more files.

The first expression defines the name of the desired parameter. With the next command a NetCdf file can be opened. This function is from the package "ncdf4" and is only available after installing and loading the package. Afterwards the degrees of longitude and latitude as well as the time, respectively hours, are captured. For each of these values, a dimension is assigned. For the longitude this would be "`nlon <- dim(longitude)`".

Afterwards a matrix is generated. This puts all the values of the wind vector "u50m" into the matrix. This matrix is then transformed into a vector and merged to a matrix with the wanted order:

```
u50m.mat <- matrix(u50m.vec.long, nrow = nlon * nlat, ncol = ntime)
```

The result is a matrix, which shows hours in the columns, from 0 to 23, and the number of grid points in the rows. But because the values of the required parameter, in this case "u50m", have to be assigned to a coordinate, the next step is to generate a data frame with all the possible coordinates – pairs of the longitudinal and latitudinal degrees. Therefore the possible combinations of longitudinal and latitudinal degrees are expanded with "`lonlat <- expand.grid(longitude,latitude)`". In the case of the Austrian MERRA-dataset, this results in 120 points.

The next expression calculates the distance of all those coordinates to a desired initial point – in particular the wind farms. This is discussed in chapter 2.1.5.2.

With

```
dfu50m <- (data.frame(cbind(dista,lonlat, u50m.mat)))
```

a new data frame is generated. It is ordered in ascending distance to the initial point and after renaming the names of rows and columns, the final data frame is available. Figure 5 shows an extract of such a data frame. In this extract, the values of the first 3 hours of the parameter "u50m" for the closest points to the wind farm, as well as the distance and the coordinates are shown.

The last part of the developed function closes the file and returns the desired data-frame. For the remaining 6 parameters, there have also been developed similar reading functions. They can be used for New Zealand and Austria likewise.

A reading function for dates and times was developed which consists of the following program code:

```
datum <- function(ncname) {
  ncfile <- nc_open(ncname)
  h <- ncvar_get(ncfile, "time")
  d <- unlist(strsplit(ncfile$dim$time$units, " "))
  date <- rep(d[3],24)
  dh <- paste(date,h)
  x <- as.POSIXct(strptime(dh, format="%Y-%m-%d %H",tz="UTC"))
  nc_close(ncfile)
  return(x)
}
```

R Program-Code 2: Reading function for timestamps of MERRA-files, source: own code

This function reads the hours and date of a file and returns a "POSIXct-vector" – which is a class in RStudio. After the file was opened, the hours ("h") and date ("d") are assigned to a variable. Since one day has 24 hours, the date is reproduced 24 times and afterwards these 2 vectors are joined ("dh"). To convert to the required format, that is later used to merge the different data-frames and match the different time zones, the "as.POSIXct" function of RStudio is applied to the vector "dh". Afterwards the file is closed and the function returns the dates and times in a useful format.

#### 3.2.3.2. DISTANCE FROM MERRA GRID POINTS TO THE WIND FARMS

Like mentioned in chapter 2.1.3.1 the following term in the reading function

```
dista <- 6378.388 * acos(sin(lat) * sin(lonlat[,2]) + cos(lat) *
              cos(lonlat[,2]) * cos(lonlat[,1]-long))
```

calculates the distance from an initial point – the examined wind farms – to the points from the dataset. The 2 variables "`lat`" and "`long`" represent the coordinates of the wind farms. It is necessary to transform the degrees of longitude and latitude from both, the wind farm and the MERRA grid points, into radians. This is what the term "`lonlat <- lonlat*rad`" does in the reading function, whereby "*`rad`*" calculates from degrees to radians by means of "$\pi/180$". For the coordinates of the wind farm this was done separately, since they are not captured in the reading function.

With equation (3), the central angle between 2 points on a sphere can be calculated. It is called the spherical law of cosines:

$$\Delta\sigma = arccos * (sin\emptyset_1 * sin\emptyset_2 + cos\emptyset_1 * cos\emptyset_2 * \cos(\Delta\lambda)) \qquad (3)$$

With:

$\Delta\sigma$ .......... central angle between 2 points

$\emptyset_1$........... geographical latitude of point 1

$\emptyset_2$........... geographical latitude of point 2

$\Delta\lambda$ .......... difference between geographical longitude of point 2 and 1

To get the distance from the wind farms to the grids of the MERRA-dataset in kilometres, the central angle between 2 points is multiplied with the radius of the earth, since the angle is calculated in radians. The radius of the earth is assumed with 6378.388km.

Because the term is used in the reading function, all distances from all points in the dataset to a given point – the wind farms – are calculated and then ordered in ascending distance with

```
dfu50m <- dfu50m[ order(dfu50m[,1]), ]
```

For the Austrian wind farm, the exact location cannot be published due to data protection requirements, as mentioned before. For New Zealand, the coordinates mentioned in chapter 2.1.2 are used for calculating the distances. The following table shows the distances in kilometres and the geographical coordinates in degrees from the 3 closest points from the MERRA grid points to the wind farms as well as the values for the parameter “u50m” for the first 3 hours of a random day.

| MERRA Point | Distance in km from wind farm | Longitude | Latitude | Wind velocities in m/s for 3 hours | | |
|---|---|---|---|---|---|---|
| | | | | 1 | 2 | 3 |
| | | | | | | |
| **White Hill Wind Farm** | | −45.7525°N 168.271667°E | | | | |
| 1 | 34.68 | 168.00 | -46.00 | 3.51 | 3.80 | 4.19 |
| 2 | 35.18 | 168.00 | -45.50 | 2.94 | 3.45 | 3.94 |
| 3 | 41.19 | 168.67 | -46.00 | 3.38 | 3.77 | 4.13 |
| | | | | | | |
| **Mahinerangi Wind Farm** | | $-45.760556°N$ $169.905°E$ | | | | |
| 1 | 27.65 | 170.00 | -46.00 | 4.82 | 5.29 | 5.64 |
| 2 | 29.93 | 170.00 | -45.50 | 4.36 | 4.64 | 4.84 |
| 3 | 51.70 | 169.33 | -46.00 | 3.96 | 4.36 | 4.67 |
| | | | | | | |
| **Te Apiti Wind Farm** | | $-40.296111°N$ $175.808333°E$ | | | | |
| 1 | 27.92 | 176.00 | -40.50 | 6.52 | 7.74 | 9.51 |
| 2 | 36.78 | 176.00 | -40.00 | 9.07 | 10.07 | 11.08 |
| 3 | 46.23 | 175.33 | -40.50 | 8.29 | 8.95 | 9.59 |

Table 1: Distance in kilometres from 3 closest MERRA grid points to the wind farms, including their coordinates and the wind velocities for the first 3 hours from a random day, source: own table

From table 1 the 3 closest points to each wind farm can be derived. The closest point is used in the further simulation of wind speeds or respectively electricity generation. It also shows, besides the distance in kilometres from the wind farm, that the wind speeds for different MERRA-points are quite different. For example, the second closest point in Te Apiti shows much higher wind speeds in all 3 hours than the closest point. This has, of course, a significant impact on the simulation and is discussed later in chapter 5. Table 1 also shows the spatial resolution of the MERRA-grid, since all longitudinal degrees have a 0.66° interval and the latitudinal degrees show a 0.5° degree interval.

### 3.2.3.3. Data Frame with all MERRA Files

Table 2 shows only the extract of one day, respectively one MERRA file, for one parameter. The next step is to generate data frames with all files and all parameters. For each parameter an own data frame is generated.

At first, all files are put together in one list. This list contains all file names of the MERRA dataset. For New Zealand this is done with the command:

```
NZfiles <- list.files(pattern = "*.nc")
```

Afterwards the function "`lapply`" is used to apply the reading functions to all files at once. This generates a list with 3652 data frames. This means each list element contains 1 table for one location for the complete period. Exemplary for the White Hill wind farm and the parameter "u50m" this would be:

```
Listu50mWH <- lapply(NZfiles, readu50m)
```

Due to the fact that the distance calculation is contained in the reading functions, the point with the shortest distance is always placed in the first row of each data frame within the list. This means that the first row of each data frame is extracted and then merged together in a new data frame. This is, again, only an example for 1 parameter. Here it is shown, again exemplary, for the White Hill wind farm and the parameter "u50m":

```
NNWHu50m <- unlist(sapply(Listu50mWH, function(d) d[1,4:27]))
```

The new data-frame now contains all values for 1 parameter for each hour and all days. Since the coordinates are no longer needed, they are dropped out of the data frame – only the columns 4 to 27, the values for the parameter, are used.

The same procedure is now done with the remaining 6 parameters, whereby equation (1) is applied to the data frames with the wind vectors $u$ and $v$. The result is the wind velocity and a reduction from 7 to 4 data frames. So there is 1 data frame for the displacement height and 3 for the wind velocities at different heights – 2m, 10m and 50m.

For the timestamps the function "datum" is used. It is applied to all files and returns a list with all dates and times for all MERRA-files. After applying the function "unlist" to the date list, a vector with the times and dates is generated. The problem, that this vector does not contain the "POSIXct" class values but instead numerical values, is solved by applying the following command:

```
MTZ <- as.POSIXct(NNdate, origin = "1970-01-01 00:00:00 UTC",
                  tz="Etc/Universal")
```

This command is necessary to transform the numerical values back to the "POSIXct" class. The used time zone is the same like the MERRA-files time zone – UTC (Universal Time Code).

All data-frames, the 4 for the parameters as well as the one for the timestamps, are then put together into 1 data-frame. It contains all hourly values for all parameters for the closest MERRA-point and the corresponding timestamps.

### 3.2.4. Time Zones

As mentioned before, the MERRA data is in UTC – Universal Time Code (same as GMT – Greenwich Mean Time). It has no time shifts during a year. The time for the MERRA data is always in half past hours, like mentioned in chapter 2.2.1., due to the fact that it is an averaged value for 1 hour. This fits quite well with the production data, because it always contains the production of a full hour, e.g. from 01:00 to 02:00, hence the average value within 1 hour should reflect the production to some extent[2].

The fact that the production data includes time shifts due to daylight saving times, which vary from year to year, has to be addressed.

The production data for Austria is in UTC+1 and there is a daylight saving time from around end of March to end of October, which means UTC+2 (i.e. CET – Central European Time, and respectively CEST – Central European Summer Time).

In New Zealand NZST – New Zealand Standard Time, which is UTC+12, and respectively NZDT – New Zealand Daylight Saving Time which is UTC+13 are used. NZDT applies from around end of September to beginning of April (southern hemisphere).

Hence the date and time for the real production data was read from the production files and then put into a data-frame – 1 column for the timestamps, 1 column for the production. The timestamps in this data-frame are also of the class "POSIXct" which contains also the time zone. Therefore it is possible to merge the two data-frames by their timestamps. This means, that RStudio recognizes the different time zones and calculates the time shifts by itself. The command

```
WHDF <- merge(WH, MWH, by.x = "NZdate", by.y = "MD")
```

generates a data-frame that is merged by its timestamps (exemplary for the White Hill wind farm). Therefore the production data and the MERRA-data are matched now on temporal scale and this final data-frame that can provide the information needed for the simulation. It contains the production and the timestamps for the local times of the wind farms plus the matched wind speeds and the displacement height from the MERRA-files.

[2] It has to be mentioned that averaged values cannot reflect the production exactly due to the non-linear characteristic of the power curve. A change in wind speed during an hour is taken into account in the real production data but not for simulated production.

As mentioned in the chapters 2.1.1, 2.1.2.1, 2.1.2.2 and 2.1.2.3, some data had to be removed, because the production data was somehow erroneous. This can easily be done by deselecting the rows of the data-frame that are not examined, e.g.:

```
WHDF <- WHDF[c(3001:30999,32751:51122),]
```

This code shows exemplary for the White Hill wind farm how to handle the removal of data. The first 3000 hours as well as the production hours 31000 to 32750 are dropped. Altogether there are 4 data-frames, one for each examined wind farm.

### 3.2.5. Extrapolating Wind Velocity – Power Law

Basically there were 2 possibilities to extrapolate the wind velocities from a given height (2m, 10m, 50m) to the hub heights of the turbines – the empirically derived power law and the logarithmic law. For the logarithmic law there would have been 2 elements of uncertainty because the roughness coefficient as well as the friction coefficient would have been needed to calculate the wind speeds for the wanted heights. Since the MERRA data contains wind speeds for 3 different heights it is quite simple to calculate the wind shear coefficient α.

As stated in [33], the value of the wind shear coefficient can be at least tripled during one day on the same location because it depends on many variables like atmospheric stability, wind speed, temperature, height, land features, and other factors. It has no physical foundation and is an engineering formula that more or less is an expression of the (in-) stability of the atmosphere. Notable is also that the power law is only used to describe wind profiles in the lower atmosphere up to around 100m. Equation (3) shows how the wind shear coefficient alpha can be calculated:

$$\propto = \frac{\ln(v_2) - \ln\,(v_1)}{\ln(h_2) - \ln\,(h_1)} \qquad (3)$$

With:

$v_2$........... wind speed at 50m above surface

$v_1$........... wind speed at 10m above displacement height

$h_2$........... 50m above surface

$h_1$........... 10m above displacement height

α...................... wind shear coefficient

After applying equation (3) to the final data frames mentioned in the previous chapter, respectively to 2 different wind speeds, the result is the wind shear coefficient for every single hour. In other words, each wind speed value can be linked with a shear coefficient value. Since the wind speeds in the MERRA-data set used for calculating the shear coefficient is 50m above surface and 10m above displacement height, the displacement height has to be added to the height $h_1$.

After rearranging equation (3) to equation (4) it is obvious that with the given data, the wind speed for the wanted height can be calculated. Equation (4) shows the power law:

$$v_2 = v_1 * \left(\frac{h_2}{h_1}\right)^{\propto} \qquad (4)$$

With:

$v_2$........... Extrapolated wind speed at hub height

$v_1$........... Wind speed at 50m above surface

$h_2$........... Hub height of the turbine

$h_1$........... 50m above surface

α...................... wind shear coefficient

By applying equation (4) the wind speeds can be extrapolated to the actual hub height of the specific wind farms.

### 3.2.6. SIMULATION

Combining the extrapolated wind speeds plus the power curves of the installed turbines, which are shown in chapter 2.2.2., the production can be simulated for each wind farm in hourly resolution. Table 2 shows the power (kW) of the turbines in dependency of the wind speed (m/s). The values in Table 2 are only shown up to 22m/s due to the fact that the maximum wind speed that occurs in the MERRA-dataset, respectively in the extrapolated wind speeds, is 21.24m/s (Te Apiti).

In order to use the values as a function in RStudio, 2 vectors ("wind" and "power") are generated for each turbine. The first vector includes the wind speeds, the second the power output. Afterwards the following command is applied to those vectors:

```
f <- approxfun(wind, power)
```

The command "approxfun" returns a function that linearly interpolates the values of the vectors. For each turbine a separate function is generated and these functions can afterwards be plotted. The results of plotting the power curves are displayed in Figure 4.

| Wind speed (m/s) | Austria<br>Enercon E70-E4 | White Hill<br>Vestas V80-2MW | Mahinerangi<br>Vestas V90-3MW | Te Apiti<br>Vestas/Micon<br>NM72-1650 |
|---|---|---|---|---|
| **0** | 0 | 0 | 0 | 0 |
| **4** | 56 | 66.3 | 75 | 0 |
| **5** | 127 | 152 | 187 | 79 |
| **6** | 240 | 280 | 348 | 204 |
| **7** | 400 | 457 | 574 | 370 |
| **8** | 626 | 690 | 875 | 576 |
| **9** | 892 | 978 | 1257 | 808 |
| **10** | 1223 | 1296 | 1688 | 1067 |
| **11** | 1590 | 1598 | 2118 | 1308 |
| **12** | 1830 | 1818 | 2514 | 1507 |
| **13** | 1950 | 1935 | 2817 | 1610 |
| **14** | 2050 | 1980 | 2958 | 1645 |
| **15** | 2050 | 1995 | 2994 | 1650 |
| **16** | 2050 | 1999 | 2999 | 1650 |
| **17** | 2050 | 2000 | 3000 | 1650 |
| **18** | 2050 | 2000 | 3000 | 1650 |
| **19** | 2050 | 2000 | 3000 | 1650 |
| **20** | 2050 | 2000 | 3000 | 1650 |
| **21** | 2050 | 2000 | 3000 | 1650 |
| **22** | 2050 | 2000 | 3000 | 1650 |

Table 2: Power of the turbines in dependency of the wind speed in kW, sources: [27]–[29], [30, S. 90], own figure

If the functions are applied to the extrapolated wind speeds, the power output for each hour is calculated:

```
pl1 <- sapply(v8650, FUN = "f")
```

The command “sapply” applies the approximated power curve to the extrapolated wind speeds (here exemplary shown for the wind farm in Austria) and returns the power output for 1 turbine in hourly resolution. In order to get the production of the whole wind farm, the results for each hour are multiplied with the number of the installed turbines.

## 3.3. STATISTICAL ANALYSIS OF SIMULATED DATA

In order to validate the MERRA-data, the simulated power output and the real power output are compared and analysed for different time spans. Several statistical parameters are examined to figure out how well the model performs.

The simulated and the real output, in hourly resolution, plus the dates are merged in a new data frame called “`mm`”, exemplary for Austria. Since the dates are in class “POSIXct”, different time collections can be extracted by using the “`format`” function of RStudio as follows:

```
ZY<-format(mm[,1],"%Y")
ZYm<-format(mm[,1],"%Y%m")
ZYmd<-format(mm[,1],"%Y%m%d")

zag_y<-aggregate(mm[,2:3],by=list(ZY),sum)
zag_mon<-aggregate(mm[,2:3],by=list(ZYm),sum)
zag_day<-aggregate(mm[,2:3],by=list(ZYmd),sum)
```

R-Program Code 3: Assigning hourly production to different time units, source: own Code

Hours, days, months and years can be identified and collected in character strings. It is important to mention, that the “`format`” function works like a “filter” that picks out parts of the date-string which can afterwards be used to assign the hourly values to the created date-strings. For example the first term creates a string that consists only of the different years. This means each hourly value is assigned to the year in which it occurs. The second term generates a string that combines years and months and hence each value is assigned to the year plus the month in which it occurs. The third does the same for years, months and days.

Afterwards these date-strings can be used to aggregate the power output with the function “`aggregate`”. To sum up the production for each single year, the first term is used. It creates a data frame with the hourly aggregated power output for each year. The second term aggregates the hourly production for each month and the third term aggregates the production for all single days. R-Program Code 2 shows exemplarily the processing for the wind farm is Austria. In this case there are in total 44.481 hours which cover 6 years, 62 months or 1854 days.

The results are prepared in 2 different parts. The first part presents and analyses the individual wind farms, the second one compares the wind farms with regard to the differences and possible common biases and errors.

For the Austrian wind farm only the statistical values for a single turbine can be shown due to data protection requirements.

The correlation is given for several temporal resolutions – hourly, daily, monthly, seasonally and annually – and all wind farms. The correlation coefficient between real and simulated production is in most cases the Pearson correlation coefficient, since in most of the cases the sample size is high enough and the assumption that the relation between the variables is linear can be made. Also the distribution is not that highly skewed and therefore, the Pearson correlation coefficient should be the first choice. Like [34] mentions "for moderately skewed distribution ... Pearson´s correlation coefficient remains the most powerful". Since the data used here is not being considered as normally distributed, the Pearson correlation coefficient is usually not the first choice in general. But [34] also shows that the Pearson correlation coefficient can be "successfully used for analysis of continuous non-normally distributed data". Therefore the only reason for using Spearman´s correlation coefficient instead of Pearson was the sample size. If the sample size is smaller than 25 the Spearman correlation is given additionally, since in [35] it is recommended that Pearson should only be used if the sample size is 25 or higher.

The 95%-confidence interval´s for the correlation are calculated with the help of the R-package "psychometric" and the command "CIr". By means of the correlation coefficient, the sample sizes, which were used for calculating the correlation coefficients, and an alpha level of 0.05 the confidence intervals can be calculated (e.g. the hourly confidence interval for the White Hill wind farm was calculated by means of a correlation coefficient of 0.7, a sample size of 61696 production hours and an alpha level of 0.05). The confidence intervals are interpreted as followed. For many thousands of samples with the sample size of the calculated correlation coefficients, the correlation coefficient of those samples will be within the confidence intervals for a proportion of 1-α (for a 95% confidence interval α=0.05) [36].

Besides the correlations and their confidence intervals, a common summary is given which includes the standard deviation, minimum, 1st quartile, median, mean, 3rd quartile and maximum for both, simulated and real production as well as the difference of real and simulated production and for a single turbine.

For each wind farm a histogram is shown that reflects the production frequencies for different power output intervals for real and simulated production. Also, the aggregated monthly simulated and real production is shown in a plot for each wind farm. 2 extra plots are presented for the Austrian wind farm that show (1) the real and simulated production for 70 randomly selected hours and (2) a 2d-contourplot that demonstrates the relationship between simulated and real production.

The second part of the results compares the 4 wind farms regarding their statistical parameters. Therefore the values of real and simulated production are normalized by the capacity of the particular wind farm and then the difference between real and simulated production as well as the real and simulated production is presented in one table. This table includes, besides the above mentioned summary of statistical parameters the correlation values and their 95% confidence intervals for all examined temporal resolutions and wind farms. The last figure of this part shows 8 box plots, 2 for each wind farm whereby 1 reflects the real and the other the simulated production.

# 4. RESULTS

## 4.1. AUSTRIAN WIND FARM

Table 3 shows an overview of the examined statistical parameters for 1 single turbine for the Austrian wind farm. Due to data protection requirements, the full parameters cannot be provided for this wind farm. It shows rounded integers, except the correlation and the coefficient of determination values, for a single turbine. The values of the single turbine are the simulated production divided by the number of installed turbines. Because of the data constraints, the focus for this wind farms is on the correlation coefficients. The mean of the simulated production is about 86.98% of the real production. The data covers a time span of 44.481 hours (1.854 days, 62 months, 21 quarters, 6 years). The abbreviations for the temporal resolutions are: h. = hours, d. = days, m. = months, q. = quarters, y. = years.

<table>
<tr><th colspan="7">Austria – 44.481 production hours, 2.050kW/turbine</th></tr>
<tr><th>Parameter</th><th>Real Prod. (kWh)</th><th>Simulated Prod. (kWh)</th><th colspan="2">Δ Real-Simulated (kWh)</th><th colspan="2">Simulated Prod. for 1 turbine (kWh)</th></tr>
<tr><td><b>Minimum</b></td><td></td><td></td><td colspan="2"></td><td colspan="2">0</td></tr>
<tr><td><b>1st Quartile</b></td><td></td><td></td><td colspan="2"></td><td colspan="2">49</td></tr>
<tr><td><b>Median</b></td><td></td><td></td><td colspan="2"></td><td colspan="2">178</td></tr>
<tr><td><b>Mean</b></td><td></td><td><b>Data Protection</b></td><td colspan="2"></td><td colspan="2">361</td></tr>
<tr><td><b>3rd Quartile</b></td><td></td><td></td><td colspan="2"></td><td colspan="2">487</td></tr>
<tr><td><b>Maximum</b></td><td></td><td></td><td colspan="2"></td><td colspan="2">2050</td></tr>
<tr><td><b>Standard Deviation</b></td><td></td><td></td><td colspan="2"></td><td colspan="2">455</td></tr>
<tr><td><b>R – Pearson Correlation Coefficient</b></td><td>Hourly<br>(44481 h.)<br>0.75</td><td>Daily<br>(1854 d.)<br>0.85</td><td>Monthly<br>(62 m.)<br>0.94</td><td colspan="2">Seasonally<br>(21 q.)<br>0.96</td><td>Annually<br>(6 y.)<br>0.99</td></tr>
<tr><td><b>R² (coefficient of determination)</b></td><td>Hourly<br>0.56</td><td>Daily<br>0.72</td><td>Monthly<br>0.89</td><td colspan="2">Seasonally<br>0.93</td><td>Annually<br>0.99</td></tr>
</table>

Table 3: Statistical Parameters for the wind farm in Austria, source: own table

The (Pearson) correlation increases with a decreasing temporal resolution, from about 75% for hourly production up to 99% for the annual production. A coefficient of determination ($R^2$) of about 0.56 reveals that more than half of the real production is explained by the simulated production on hourly basis. On a monthly basis, which includes 62 months, $R^2$ is 0.89.

Figure 8 represents a histogram for simulated and real production on hourly based values for a single turbine. Each bar contains the values for a range of 100kWh. Red colour marks a surplus in real production and light-blue represents a surplus of the simulated production. From 0 to 100kWh the real production contains more values, which could be a consequence of a production stop caused by the operator[3] and not mainly due to lack of wind. For the bins between 100kWh and 1.200kWh the simulated production has a slightly higher frequency than the real production. The frequency decreases with the increase of the power. From 1.300kWh upwards the real production has a continuous higher frequency than the simulated production. This explains quite well the reason for the higher total output of the real compared to the simulated production which was mentioned above. Worth to mention is that the maxima (single highest output) of both are quite the same (see Table 3), but this is by far not the case for the frequency of them, since the real production has a more than twice as high frequency for a production of 2.000+ kWh compared to the simulated production.

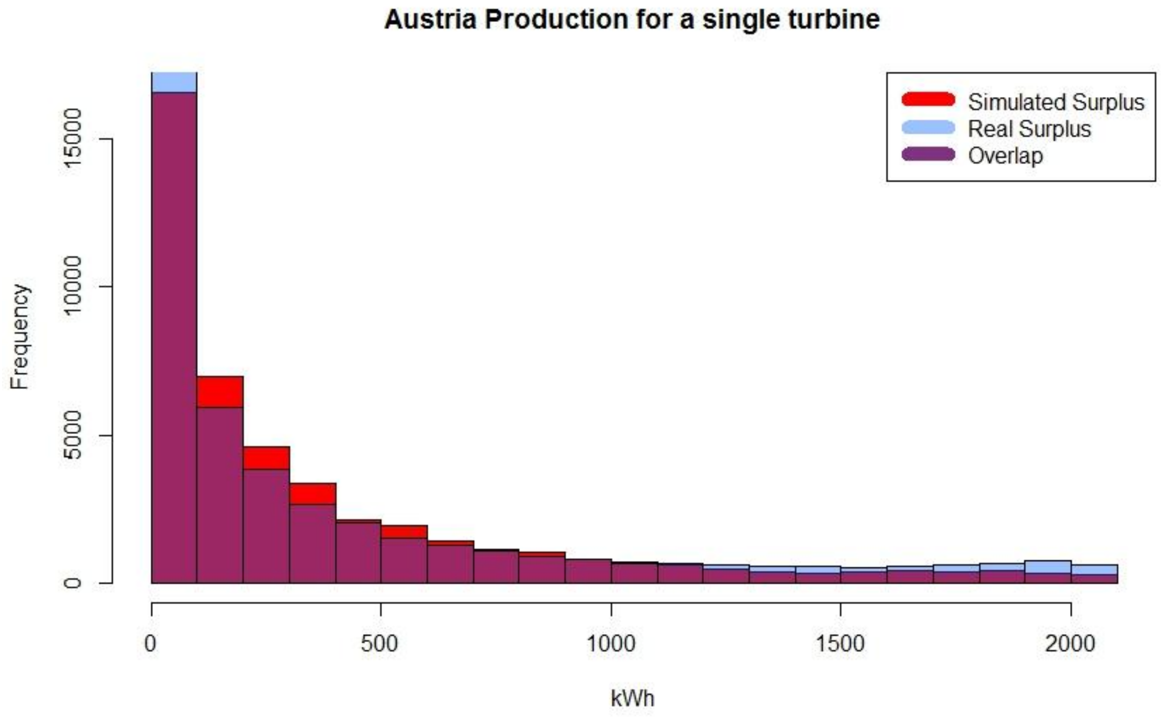


Figure 8: Histogram of hourly production data for Austria, source: own figure

[3] Possible reasons could be maintenance processes or a necessary shut down of the production due to several reasons, e.g. overload of the grid.

In figure 9 the production for 70 sequential hours, (from hour 25.000 to hour 25.070), is shown. It can clearly be seen that the simulated production has a much smoother behaviour and has a lower production than the real production for that period. Hardly any peaks can be seen in the simulated production in opposite to the real production. And for higher wind speeds, respectively power output, the real values are continuously higher than the simulated ones. For lower wind speeds a different picture can be observed – the simulated production is most of the times slightly higher compared to the real one. The huge difference in the middle (hour 28 – 35) plus the higher simulated production for lower power output confirms the information given by the histogram in figure 8.

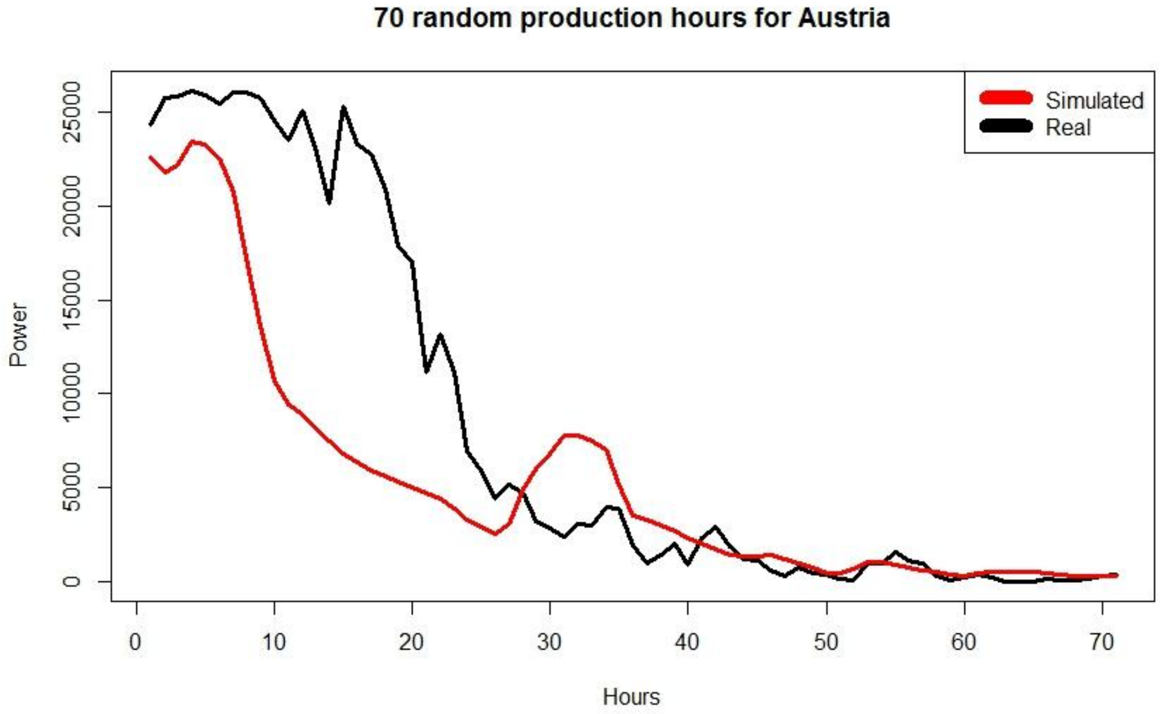


Figure 9: 70 hours (sequential, from hour 25.000 to hour 25.70) for real and simulated production
Source: own figure

A look on the monthly production shows a quite similar picture, which can be seen in figure 10. The production data was aggregated to single months and then plotted in the same figure. The simulated production is nearly continuously lower than the real production. Still, a quite good correlation of around 0.94 is achieved.

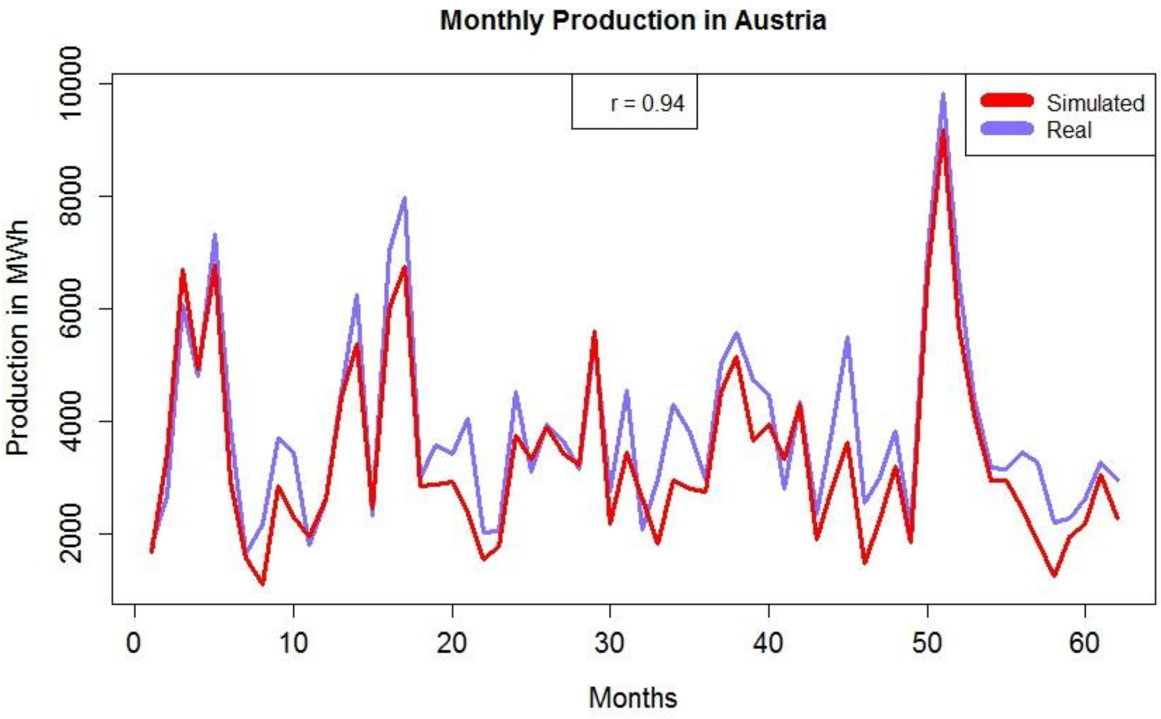


Figure 10: Monthly Production in Austria, source: own figure

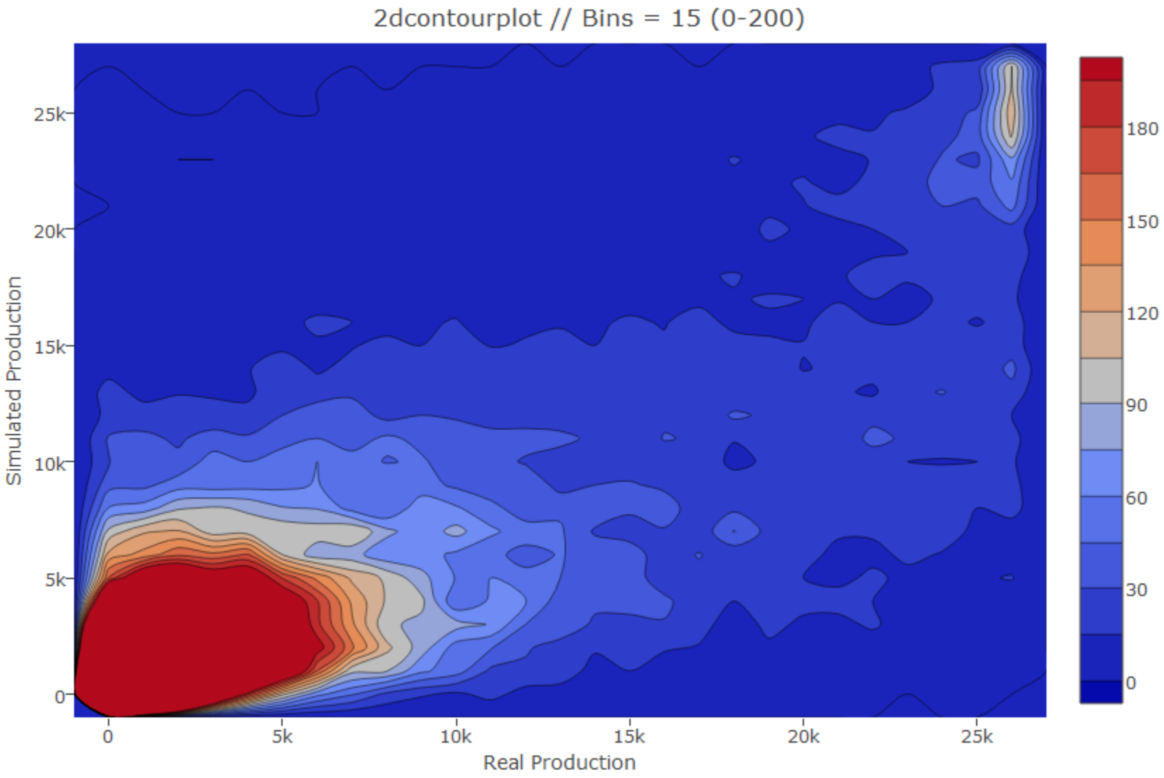


Figure 11: 2-d contour plot for hourly production in Austria, source: own figure

The 2d-contourplot in figure 11 shows the relationship between simulated and real production and their common occurrences which are represented in different colours – from dark blue to dark red – for different common occurrences. The colour scale reaches from dark blue to dark red, with different shares of red and blue. The higher the share of red, and therefore the lower the share of blue gets, the higher the common occurrences. The 3 dimensional relationship between simulation, real production and their common occurrences can be shown in 2 dimensions. The 2 hotspots of the plot can be seen for a production from 0kWh to 5.000kWh and for maximum output around 25.000kWh. The plot shows a quite significant bias. For a very low production the simulation is slightly overestimating the real production, but from 5.000kWh upwards, the simulation underestimates the real production substantially. Hardly any common occurrences can be observed for a real production lower than 15.000kWh and a simulated of 15.000kWh or higher, whereas a whole lot of common occurrences which have a real production of 15.000kWh upwards and a simulated production of less than 15.000kWh. This is also reflected in a much lower total production of the simulation.

## 4.2. WHITE HILL WIND FARM

The statistical parameters for White Hill are represented in table 4. The values for the whole wind farm (simulated and real) as well as for a single turbine are rounded integers that represent the production in kWh, except the correlation and coefficient of determination values. The production for 1 turbine is the simulated production divided by the number of turbines, which is in this case 29. A quite huge difference can be seen in the first quartile, which is 5.402kWh for the simulated and 1.360kWh for the real production, which is a difference of 4.042kWh. For higher production (represented by median, mean, the 3rd quartile or the maximum) the values become more aligned. For the 3rd quartile, the real production is even higher than the simulated production. The total sum of production is 952.524.685kWh for the real and 1.088.562.939kWh for the simulated production. In opposite to the Austrian wind farm, the simulated production is about 14.28% higher than the real one. The data covers a time span of 46.371 hours (1.934 days, 65 months, 22 quarters, 7 years).

The (Pearson) correlation increases with the increase of the temporal resolution, with the exception of seasonal against monthly correlation (however without statistical significance for the last). The seasonal correlation aggregates the production for 3 sequential months – January to March, April to June, July to September and October to December. The correlation is, in general, not that high as for the Austrian wind farm: about 70.4% for hourly resolution and 88.6% for monthly resolution. This means that only 49.6% of the real production can be described by the simulated production for hourly data, which is represented by R².

<table>
<tr><th colspan="6">White Hill – 61.696 production hours, 58.000kW capacity</th></tr>
<tr><th>Parameter</th><th>Real Prod. (kWh)</th><th>Simulated Prod. (kWh)</th><th>Δ Real-Simulated</th><th colspan="2">Simulated Prod. for 1 turbine</th></tr>
<tr><td><b>Minimum</b></td><td>0</td><td>0</td><td>0</td><td colspan="2">0</td></tr>
<tr><td><b>1st Quartile</b></td><td>1360</td><td>5402</td><td>-4042</td><td colspan="2">186</td></tr>
<tr><td><b>Median</b></td><td>12895</td><td>18390</td><td>-5495</td><td colspan="2">634</td></tr>
<tr><td><b>Mean</b></td><td>20541</td><td>23475</td><td>-2934</td><td colspan="2">809</td></tr>
<tr><td><b>3rd Quartile</b></td><td>41470</td><td>41200</td><td>270</td><td colspan="2">1421</td></tr>
<tr><td><b>Maximum</b></td><td>57500</td><td>58000</td><td>-500</td><td colspan="2">2000</td></tr>
<tr><td><b>Standard Deviation</b></td><td>19916</td><td>19702</td><td>214</td><td colspan="2">679</td></tr>
<tr><td><b>R – Pearson Correlation Coefficient</b></td><td>Hourly<br>(46371 h.)<br>0.70</td><td>Daily<br>(1934 d.)<br>0.80</td><td>Monthly<br>(65 m.)<br>0.87</td><td>Seasonally<br>(22 q.)<br>0.86</td><td>Annually<br>(7 y.)<br>0.98</td></tr>
<tr><td><b>R² (coefficient of determination)</b></td><td>Hourly<br>0.50</td><td>Daily<br>0.65</td><td>Monthly<br>0.79</td><td>Seasonally<br>0.74</td><td>Annually<br>0.96</td></tr>
</table>

Table 4: Statistical parameters for White Hill wind farm, source: own table

Figure 13 shows the histogram for hourly production data. Light blue represents a higher production frequency for the real and red a higher production frequency for the simulated production. The breaks for each bar are set with 3.000kWh. For the first bar (0 – 3.000kWh) the real production frequency is slightly higher, which is assumed to be the same reason that were mentioned for the Austrian wind farm. The frequency for a power output up to 40.000kWh is continuously higher for the simulated compared to the real data, whereby the differences decrease with the increase of the power output. From 40.000kWh up to about 55.000kWh the frequency is higher for the real production, which is quite the same like for the Austrian wind farm. The extraordinary higher frequency for a power output above 55.000kWh for the simulated production could be a consequence of production restriction. In figure 4 – electricity production for White Hill wind farm – 1 single peak is visible for production hour 18.975. This peak value of 57.500kWh is only reached once in the whole dataset. The second highest value is 56.200kWh, the third highest 56.150kWh. This might support the assumption that the operator is restricting production in the wind farm due to e.g. limited interconnection capacities.

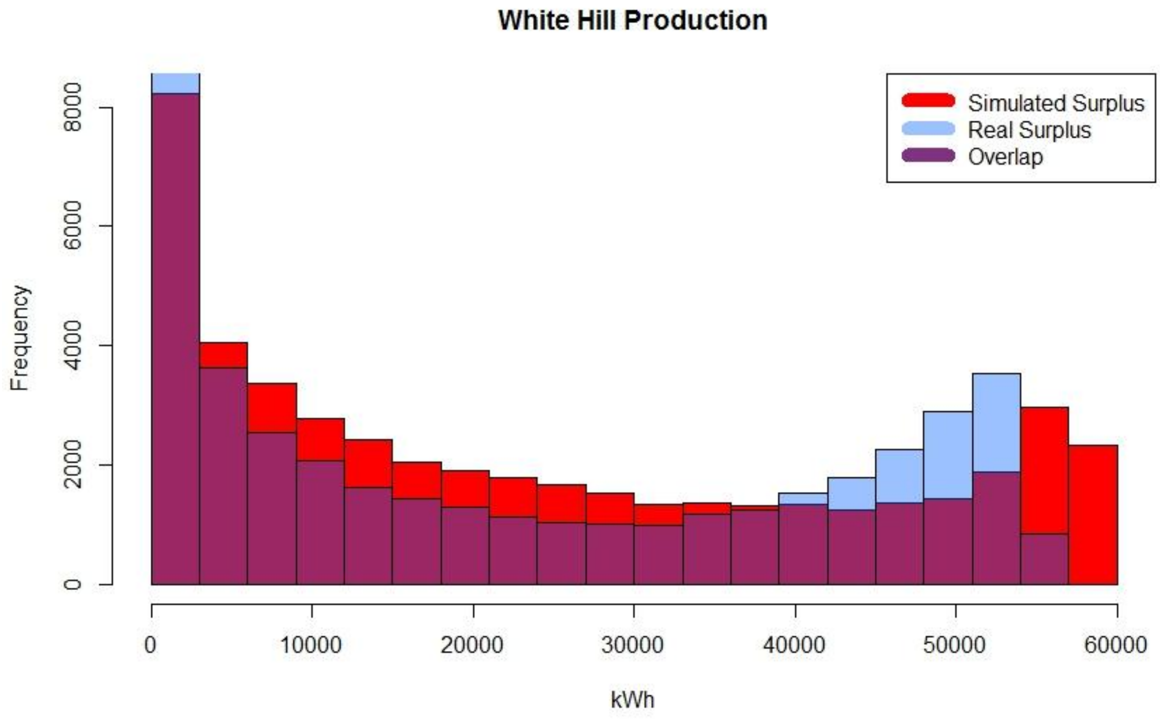


Figure 12: Histogram for hourly production data for White Hill, source: own figure

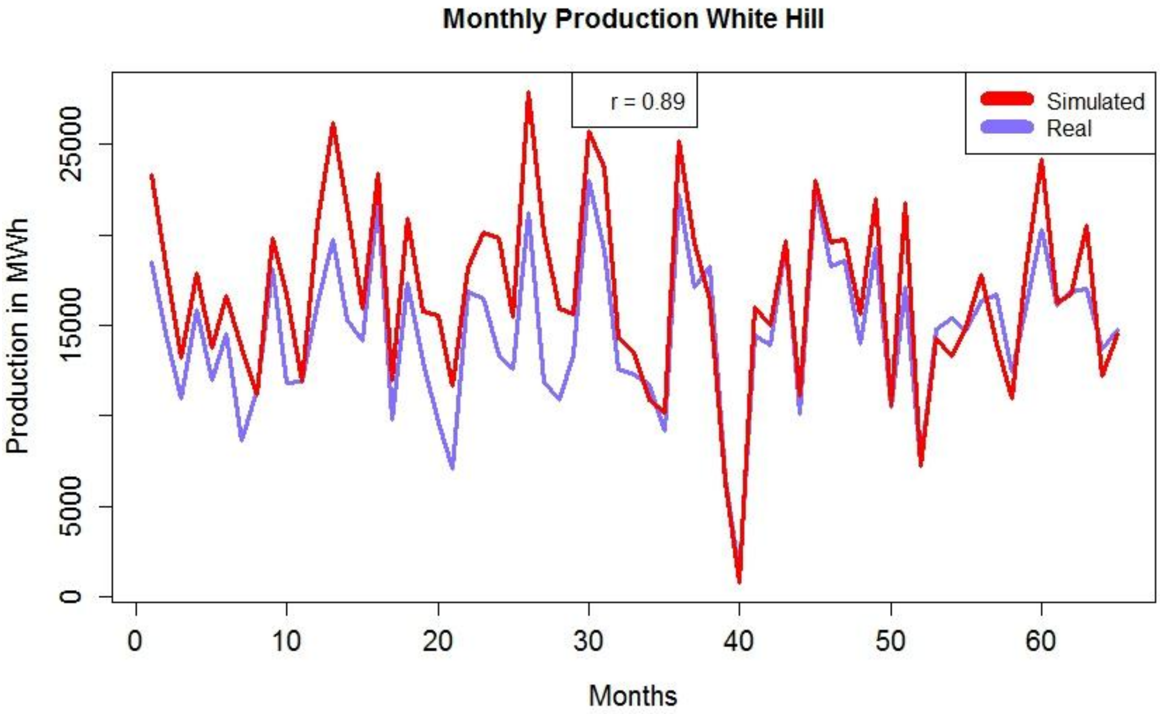


Figure 13: Monthly production for White Hill, source: own figure

The monthly production is shown in figure 14. The production data was aggregated to single months and plotted against each other. A bit of a reversed situation compared to the

Austrian wind farm can be observed here. The real production is nearly continuously lower than the simulated one, except for a few single months in which the production was lower in total. This could also reflect production restrictions imposed by the operator.

## 4.3. MAHINERANGI WIND FARM

Table 5 shows the statistical parameters for the Mahinerangi wind farm in rounded integers for the real and the simulated production as well as for 1 turbine, except the correlation and the coefficient of determination, which values are rounded to 2 positions behind the decimal point. The parameters are calculated for the hourly production data. The simulated production for 1 turbine is the whole simulated production divided by 12 (installed turbines in total). A quite similar result can be seen here, compared to the White Hill wind farm. The first quartile is nearly 3 times higher from the simulated compared to the real production. The harmonisation increases with higher production. Since the mean value is higher for the simulated production, it is clear that the total sum of production is higher too. The real production had an output of 203.011.205kWh and the simulated output would have been 211.185.422kWh. The overproduction of about 4% seems to be a quite good result. The data covers a time span of 17.473 hours (730 days, 25 months, 9 quarter, 3 years).

A look at the (Pearson) correlation coefficients shows that, while mean production is similar, the time profile of production is not captured well. For hourly production the coefficient is only 0.68, which means that only about 46% of the real production can be explained by the simulated production. Even though the correlation increases with the temporal resolution, from about 81% for daily, 92% for monthly, 98% for seasonal up to 99.9% for annual, the result could be biased because of the short observation time of only 3 years of production. Therefore the seasonal and the annual correlation are not significant.

<table>
<tr><th colspan="8">Mahinernagi – 17.473 production hours, 36.000kW capacity</th></tr>
<tr><th>Parameter</th><th>Real Prod.</th><th colspan="2">Simulated Prod.</th><th colspan="2">Δ Real-Simulated</th><th colspan="2">Simulated Prod. for 1 turbine</th></tr>
<tr><td><b>Minimum</b></td><td>0</td><td colspan="2">0</td><td colspan="2">0</td><td colspan="2">0</td></tr>
<tr><td><b>1st Quartile</b></td><td>710</td><td colspan="2">2099</td><td colspan="2">-1389</td><td colspan="2">59</td></tr>
<tr><td><b>Median</b></td><td>6850</td><td colspan="2">7926</td><td colspan="2">-1076</td><td colspan="2">571</td></tr>
<tr><td><b>Mean</b></td><td>11619</td><td colspan="2">12086</td><td colspan="2">-467</td><td colspan="2">968</td></tr>
<tr><td><b>3rd Quartile</b></td><td>21375</td><td colspan="2">20317</td><td colspan="2">1058</td><td colspan="2">1781</td></tr>
<tr><td><b>Maximum</b></td><td>36000</td><td colspan="2">36000</td><td colspan="2">0</td><td colspan="2">3000</td></tr>
<tr><td><b>Standard Deviation</b></td><td>11961</td><td colspan="2">11626</td><td colspan="2">335</td><td colspan="2">997</td></tr>
<tr><td><b>R – Pearson Correlation Coefficient</b></td><td colspan="2">Hourly<br>(17473 h.)<br>0.68</td><td>Daily<br>(730 d.)<br>0.81</td><td>Monthly<br>(25 m.)<br>0.92</td><td colspan="2">Seasonally<br>(9 q.)<br>0.98</td><td>Annually<br>(3 y.)<br>0.99</td></tr>
<tr><td><b>R² (coefficient of determination)</b></td><td colspan="2">Hourly<br>0.46</td><td>Daily<br>0.65</td><td>Monthly<br>0.85</td><td colspan="2">Seasonally<br>0.95</td><td>Annually<br>0.99</td></tr>
</table>

Table 5: Statistical parameters for Mahinerangi wind farm, source: own table

The histogram for the hourly production is shown in figure 15. A higher frequency for the real production is highlighted by a light blue colour and a higher frequency for the simulated production by red colour. The breaks for each bar are set at 2.000kWh intervals. Quite similar to the 2 wind farms presented above, the real production has a higher frequency for the first bar (0 - 2.000kWh), which is probably a consequence of the reasons mentioned in 3.1. The overestimation of the simulated production for lower wind speeds is quite the same as compared to the other 2 wind farms, as well as the underestimation for higher wind speeds, except the last bar. The reason for the much higher frequency of the simulated production for the last bar is most likely not a consequence of production restrictions, since the number of occurrences for full, or close to full production is not unusually low.

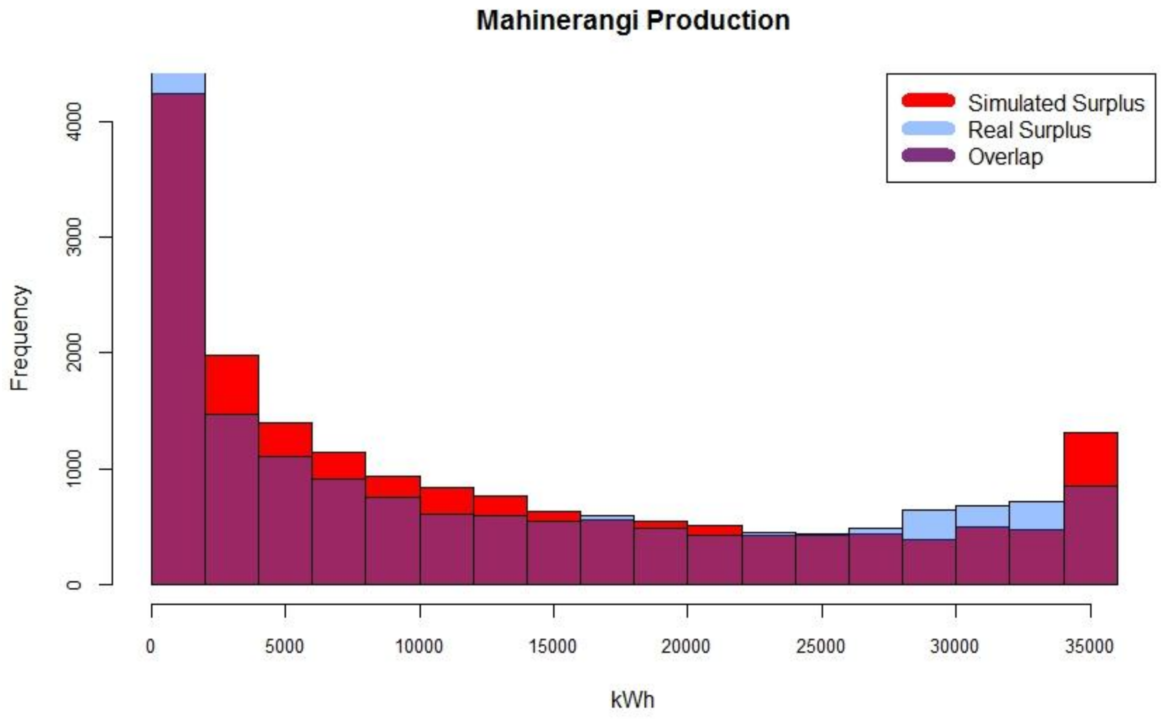


Figure 14: Histogram for hourly production data for Mahinerangi, source: own figure

The aggregated monthly production in figure 16 shows, in general, a quite good correlation, even though over- as well as underestimation of the simulated production can be observed. Due to the short time period of available production data, the relationship between monthly and annual or seasonal production, respectively correlation is not significant.

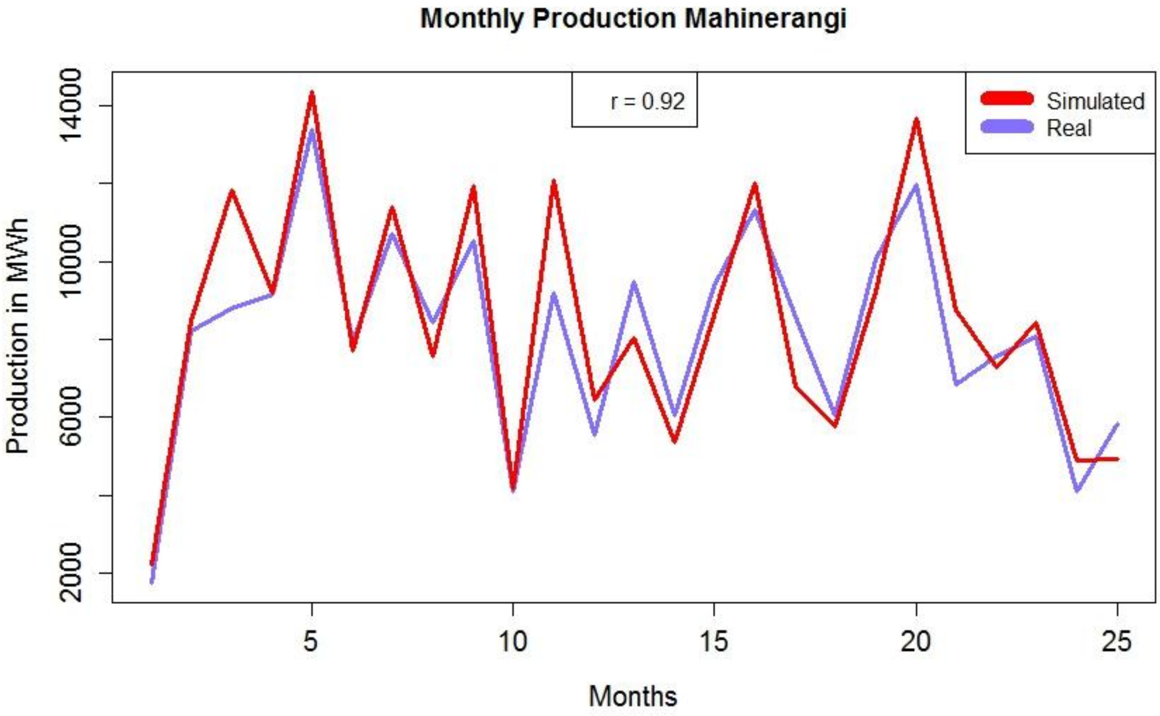


Figure 15: Monthly production Mahinerangi, source: own figure

## 4.4. TE APITI WIND FARM

Table 6 shows the statistical parameters for the Te Apiti wind farm. These results may be most significant with regard to the time span of production, since it covers more than 7 years. The values are rounded integers, with the exception of the correlation coefficient and the coefficient of determination. The values for a single turbine are the result of the simulated production divided by 55, which equals the number of turbines that are installed in the wind farm. The first quartile is slightly higher for the simulated production but it shows a quite low difference. The mean value predicts a higher total production of the real production (2.163.399.270kWh) compared to the simulated one (2.091.440.542kWh), which equals an underestimation of the simulated production of about 3.3%. The 3$^{rd}$ quartile is, like expected, slightly higher for the real than the simulated production. The data covers a time span of 61696 hours (2573 days, 86 months, 29 quarters, 9 years).

<table>
<tr><th colspan="7">Te Apiti – 61.696 production hours, 90.750kW capacity</th></tr>
<tr><th>Parameter</th><th>Real Prod.</th><th>Simulated Prod.</th><th colspan="2">Δ Real-Simulated</th><th colspan="2">Simulated Prod. for 1 turbine</th></tr>
<tr><td><b>Minimum</b></td><td>0</td><td>0</td><td colspan="2">0</td><td colspan="2">0</td></tr>
<tr><td><b>1st Quartile</b></td><td>5240</td><td>5541</td><td colspan="2">-301</td><td colspan="2">95</td></tr>
<tr><td><b>Median</b></td><td>30400</td><td>24425</td><td colspan="2">5975</td><td colspan="2">553</td></tr>
<tr><td><b>Mean</b></td><td>35065</td><td>33899</td><td colspan="2">1166</td><td colspan="2">638</td></tr>
<tr><td><b>3rd Quartile</b></td><td>63266</td><td>59709</td><td colspan="2">3557</td><td colspan="2">1150</td></tr>
<tr><td><b>Maximum</b></td><td>90470</td><td>90750</td><td colspan="2">-280</td><td colspan="2">1645</td></tr>
<tr><td><b>Standard Deviation</b></td><td>29460</td><td>31420</td><td colspan="2">-1960</td><td colspan="2">536</td></tr>
<tr><td><b>R – Pearson Correlation Coefficient</b></td><td>Hourly<br>(61696 h.)<br>0.67</td><td>Daily<br>(2573 d.)<br>0.74</td><td>Monthly<br>(86 m.)<br>0.81</td><td colspan="2">Seasonally<br>(29 q.)<br>0.74</td><td>Annually<br>(9 y.)<br>0.99</td></tr>
<tr><td><b>R² (coefficient of determination)</b></td><td>Hourly<br>0.45</td><td>Daily<br>0.55</td><td>Monthly<br>0.66</td><td colspan="2">Seasonally<br>0.55</td><td>Annually<br>0.98</td></tr>
</table>

Table 6: Statistical parameters for Te Apiti wind farm, source: own table

The correlation shows a quite poor performance for all temporal resolutions, except for the annual observation, which however is insignificant due to the relative low amount of observations. A quite interesting result is the very low correlation of the seasonal production (73.84%), which is similar to the daily production (73.83%). For hourly resolution, the correlation coefficient is about 0.67, which means that only about 44.6% of the real production can be explained by the simulation. Even though the coefficients increase with the temporal resolution, except of the seasonal correlation, a correlation coefficient of about 0.82 for monthly production is low with regard to the long observation time span of this wind farm.

Figure 17 shows the histogram for the hourly production data for Te Apiti. Blue colour is, as before, a higher frequency of the real production and red highlights a higher frequency for simulated production. The breaks for the single bars are set in this case at 5.000kWh. The higher frequency for the real production in the first bar is the same compared to all other wind farms, assuming the same reasons. Also the overestimation of lower production and an underestimation for higher production up to a certain point can be observed in all other wind farms, even though the real production frequency reaches earlier the point in which it overtakes the simulated production, relative to the total possible output of the wind farm. The extraordinary high frequency of the simulated production for a very high power output, represented by the last 2 bars, is a phenomenon that can also be observed especially for the White Hill wind farm, even though this is the most outstanding occurrence of this phenomenon.

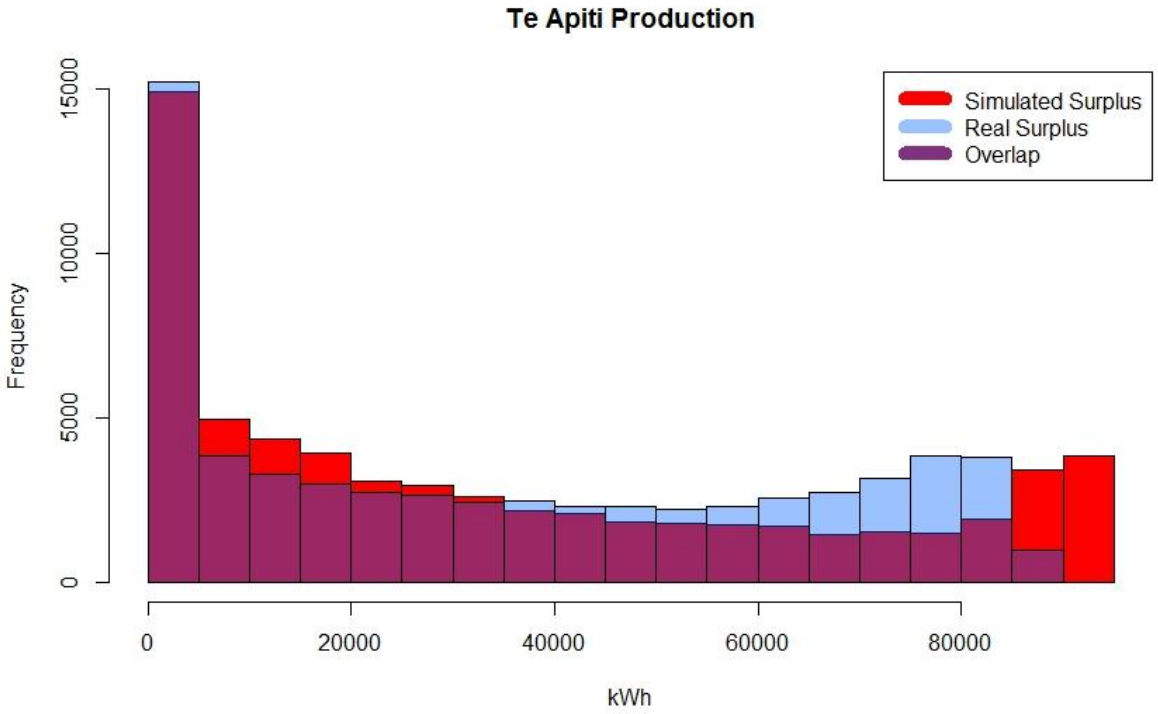


Figure 16: Histogram for hourly production data for Te Apiti, source: own figure

The reason for this high frequency is assumed to be the same like mentioned in 4.2. – it is assumed that the system operator restricts production. The highest real output for Te Apiti was 90.470kWh in production hour 28.628. The second highest output was 90.290kWh and the third highest 89.390kWh. This means, the wind farm is able to produce up to a peak from 90.470kWh. It is very unlikely that the wind conditions were that bad, that the wind farm could reach the full output only once in about 9 years. This constitutes the assumption that production restrictions were imposed.

The aggregated monthly production of Te Apiti is shown in figure 18. For some months with extraordinary high or low production (e.g. months 15 and 16, 20 to 22 or 81 and 82), the simulation represents the real production very nicely, but for others (e.g. month 54 to 56, 24 to 27) the difference between real and simulated production is quite high. In general there cannot be seen a continuous reliable compliance between real and simulated production, which is also reflected by the quite low correlation of about 82%.

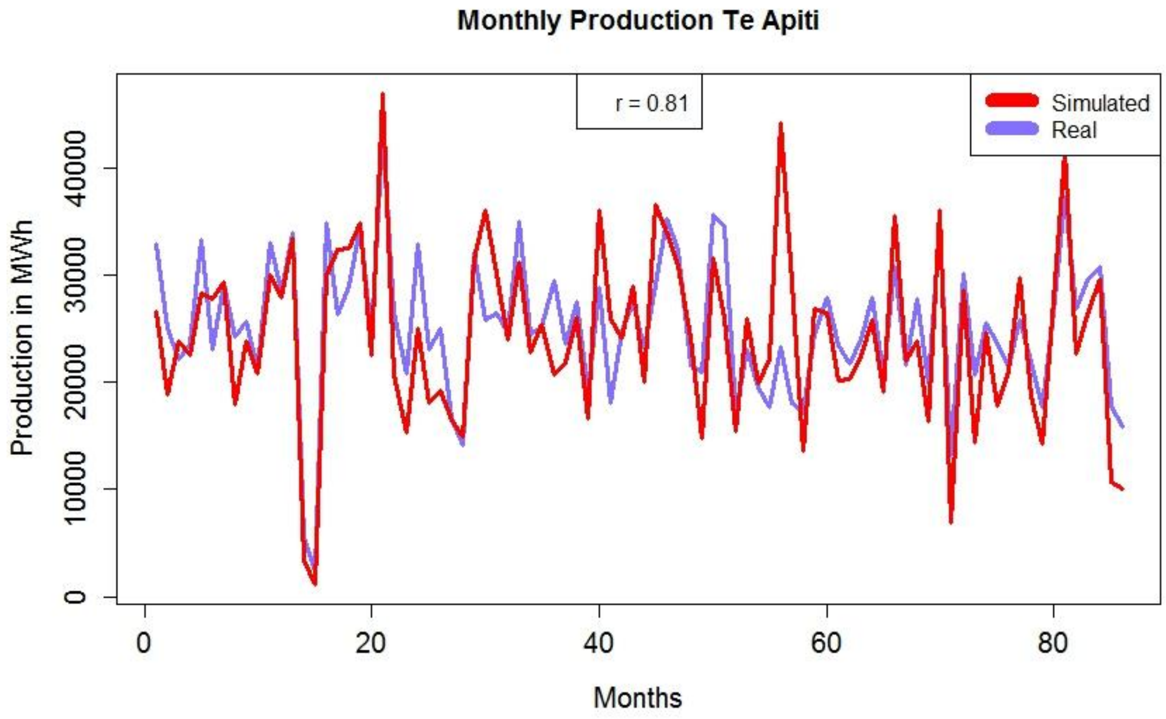


Figure 17: Monthly production for Te Apiti, source: own figure

## 4.5. WIND FARMS COMPARED

In order to compare the results of all 4 wind farms and to identify possible sources of error or biases, table 7 shows selected results of the comparison of real and simulated data. For the parameters 1$^{st}$ quartile, median, mean, 3$^{rd}$ quartile and maximum, the hourly production data, for real as well as for simulated data, were normalized by the capacity of the specific wind farm in order to make it easier to compare the results. Besides the real and the simulated hourly production, the difference between real and simulated hourly production is given for the above mentioned parameters. Due to data restriction requirements for the Austrian wind farm, the capacity cannot be stated, instead the capacity of 1 turbine plus the production hours are given. For the New Zealand wind farms, the capacity and the production hours are stated. For 2 wind farms – Mahinerangi and Te Apiti – the total amount of simulated production is quite close to the real production, whereas for the Austrian wind farm it only equals 86.98% of the real production and for White Hill it is 114.28% of the real production.

The Pearson correlation coefficients are given for all temporal resolutions. Additionally the Spearman correlation coefficient is given for samples that consist of 25 or less values, since for a small sample size the Spearman coefficient is preferable, which was mentioned in chapter 3.3.

| | | | | | | |
|---|---|---|---|---|---|---|
| | **Austria**<br>**44.481 production hours**<br>**2.050kW/turbine** | | | **White Hill**<br>**61.696 production hours**<br>**58.000kW capacity** | | |
| **Parameters** | Real | Simulated | **Δ Real-Sim.** | Real | Simulated | **Δ Real-Sim.** |
| **1st Quartile** | 0.016 | 0.024 | **-0.008** | 0.023 | 0.093 | **-0.070** |
| **Median** | 0.082 | 0.087 | **-0.004** | 0.222 | 0.317 | **-0.095** |
| **Mean** | 0.202 | 0.176 | **0.026** | 0.354 | 0.405 | **-0.051** |
| **3rd Quartile** | 0.277 | 0.238 | **0.039** | 0.715 | 0.710 | **0.005** |
| **Maximum** | 0.993 | 1.000 | **-0.007** | 0.991 | 1.000 | **-0.009** |
| **Standard Deviation** | 0.2656 | 0.2218 | **0.044** | 0.343 | 0.340 | **0.004** |
| **Total Production Ratio Simulated : Real** | **86.98%** | | | **114.28%** | | |
| | | | | | | |
| **Correlations (* Spearman coefficient for n≤25)** | | | **95% CI** | | | **95% CI** |
| **Hourly** | 0.75 | | 0.745 \| 0.754 | 0.70 | | 0.695 \| 0.705 |
| **Daily** | 0.85 | | 0.837 \| 0.862 | 0.80 | | 0.783 \| 0.815 |
| **Monthly** | 0.94 | | 0.902 \| 0.964 | 0.89 | | 0.795 \| 0.919 |
| **Seasonally** | 0.96 | 0.95* | 0.902 \| 0.984 | 0.86 | 0.79* | 0.688 \| 0.941 |
| **Annually** | 1.00 | 0.71* | 0.908 \| 1.000 | 0.98 | 0.64* | 0.866 \| 1.000 |
| | | | | | | |
| | **Mahinerangi**<br>**17.473 production hours**<br>**36.000kW capacity** | | | **Te Apiti**<br>**61.696production hours**<br>**90.750kW capacity** | | |
| **Parameters** | Real | Simulated | **Δ Real-Sim.** | Real | Simulated | **Δ Real-Sim.** |
| **1st Quartile** | 0.020 | 0.058 | **-0.039** | 0.058 | 0.061 | **-0.003** |
| **Median** | 0.190 | 0.220 | **-0.030** | 0.335 | 0.269 | **0.066** |
| **Mean** | 0.323 | 0.336 | **-0.013** | 0.386 | 0.374 | **0.013** |
| **3rd Quartile** | 0.594 | 0.564 | **0.030** | 0.697 | 0.658 | **0.039** |
| **Maximum** | 1.000 | 1.000 | **0.000** | 0.997 | 1.000 | **-0.003** |
| **Standard Deviation** | 0.332 | 0.323 | **0.009** | 0.325 | 0.346 | **-0.022** |
| **Total Production Ratio Simulated : Real** | **104.03%** | | | **96.67%** | | |
| | | | | | | |
| **Correlations (* Spearman coefficient for n≤25)** | | | **95% CI** | | | **95% CI** |
| **Hourly** | 0.68 | | 0.672 \|0.688 | 0.67 | | 0.666 \| 0.674 |
| **Daily** | 0.81 | | 0.783 \| 0.834 | 0.74 | | 0.722 \| 0.757 |
| **Monthly** | 0.92 | 0.90* | 0.825 \| 0.965 | 0.81 | | 0.722 \| 0.872 |
| **Seasonally** | 0.98 | 0.83* | 0.905 \| 1.000 | 0.74 | | 0.512 \| 0.870 |
| **Annually** | 0.95 | 1.00* | -[4] | 0.99 | 0.90* | 0.951 \| 1.000 |

Table 7: Comparison of all 4 wind farms – by capacity normalized values for upper parameters, ratio of simulated to real total production and Pearson correlation coefficients for several temporal resolutions (Spearman coefficient if n≤25), source: own table

[4] For the Mahinerangi wind farm the sample size (3 years) is too low for calculating the confidence intervals for the correlation coefficients for annual production.

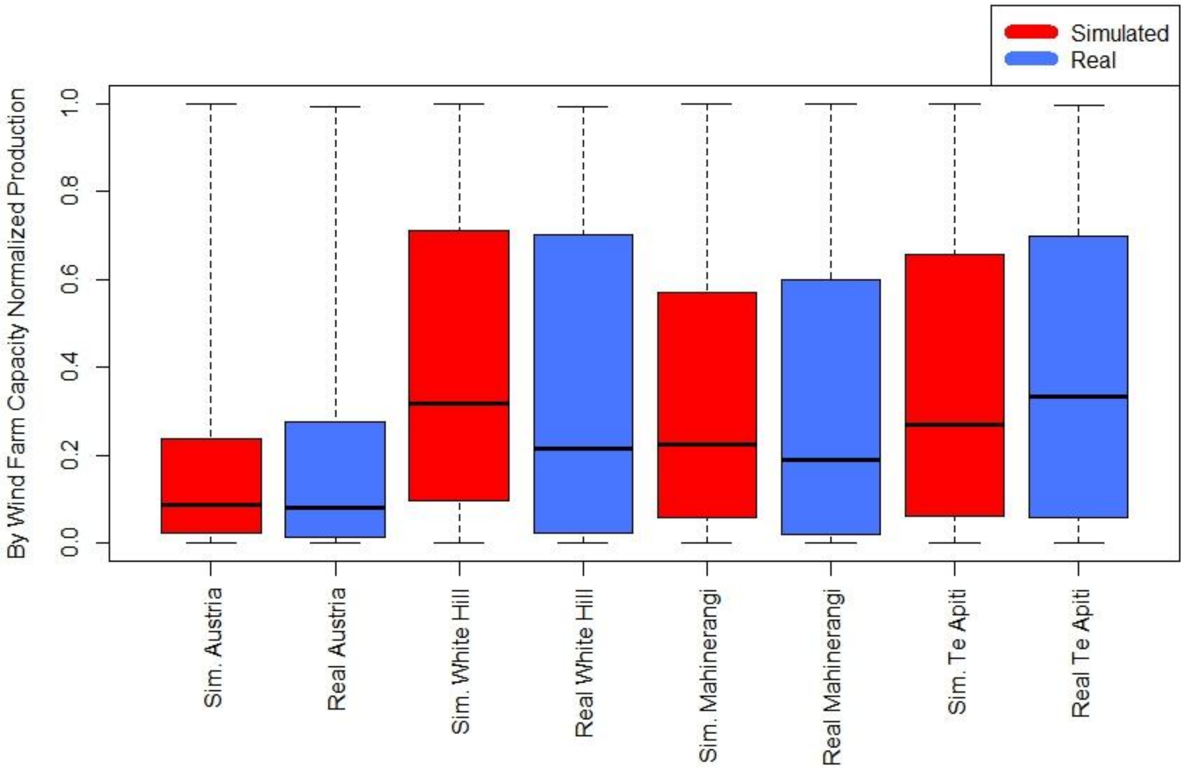


Figure 18: Box plots for all 4 wind farms, source: own figure

The box plots of figure 19 show and confirm the results that were presented in table 7 for the statistical parameters. The values used for developing the box plots are hourly production values which were normalized by the capacity of each wind farm. Red colour designates the real production and blue colour the simulated production. The box plots give a quick overview of the results mentioned before. It can clearly be seen, that for Austria the simulated production is lower and for White Hill the simulated production is higher. The distribution of the production within New Zealand is quite comparable, but it´s different for Austria. For Te Apiti and Mahinerangi the difference between simulation and real production is low, regarding the distribution.

# 5. DISCUSSION

The developed simulation model can reflect the real wind power production up to a certain point, although the results differ quite heavily for the different wind farms and temporal resolutions. Since one of the main reasons for developing this simulation model is to estimate the wind power potential for a certain area or location in a time and cost-efficient way, compared to measuring wind speeds, it has to be competitive or nearly as accurate, regarding the quality. For this purpose, a quite high temporal resolution (i.e. hourly or at least daily) is necessary. The results for hourly or daily resolution, regarding the correlation, are not that convincing and therefore the first goal cannot be achieved by this simulation model. A possible use could be to make a rough estimate of a potential wind power location without any further or precise information about potential electricity generation. For the second main reason – to use the data for large scale integration studies – lower temporal resolutions can be sufficient in order to determine the ideal integration of wind power in the power system. To determine prospective integration or respectively the share of wind power in the power grid it is not that important to know exactly the electricity generation for each single hour, rather than for longer time periods, since the power grid should be able to buffer fluctuations up to a certain point and manage different electricity generation technologies – also, geographical smoothing occurs in large integration studies. Therefore, lower temporal resolutions (i.e. monthly or seasonally) can be sufficient, which means that the developed simulation model is capable to deliver valuable results for this purpose, although the current, simple approach should still be enhanced.

First, tests with more production data at different locations should be developed. Also, an optimization of the applied approaches (i.e. interpolation to height, power curves), and bias correction may improve results significantly. A further enhancement could be an empirical derived bias correction for wind speeds (e.g. increase wind speeds frequencies of 0m/s, decrease wind speed frequencies from 4 up to about 10m/s and increase wind speed frequencies from 12m/s upwards). However, such an empirical derived optimization procedure depends on local measurement data and cannot be easily generated globally.

Probably the highest influence and therefore the major source of error is the low spatial resolution of the MERRA-data (about 50km x 50km horizontally). Table 1 shows the eastward wind speeds in 50m above surface for the 3 closest points for the 3 examined wind farms in New Zealand. The distance to the wind farms is sometimes close, although it can clearly be seen, in particular for the Te Apiti wind farm, that the wind speeds differ quite heavily for different points at the same time. Therefore, spatial interpolation procedures between the MERRA grid points may further enhance the quality of results. For that purpose, local topography should be taken into account. Further it could be useful to interpolate the wind speeds not only to a single point which is used as a representative point for the whole wind farm, but instead to the single wind power turbines, since the area of a wind farm can be quite huge (e.g. Mahinerangi 17.23km², Te Apiti 11.5km²).

Another problem with MERRA-data is the fact that the maximum wind speeds are far too low. From all examined wind farms, the highest wind speed value was 21.24m/s at a turbine height of 70m for the Te Apiti wind farm. It has to be considered improbable, that during a time span of 10 years, the highest wind speed calculated for a height of 70m is only about 24.16m/s. For the Austrian wind farm the highest observed value for 50m above surface was only 16.6m/s. Although higher wind speeds are not useful for wind power production (as production will even stop at very high wind speeds), it is obvious that the low spatial resolution of the MERRA-data causes smoothing effects and therefore lowers variance in comparison to real wind speeds. Nevertheless a study from Cannon et al. [9], which uses MERRA-data to quantify extreme wind power generation, states that "frequency and severity of extreme generation events … is found to be well reproduced by the MERRA derived time series". However, the purpose of this study was to use MERRA-data for simulating aggregated, total Great Britain wind power production and not production in a single wind farm.

A few other methodological details, like production restrictions, the simplified use of the power curves, the data removal or the assumption of active control of the turbines in wind direction had to be made. It is, however, not known if all turbines in the sample can actively control towards wind direction. Wind turbines also lose performance over time – a topic which was researched by Staffel et al. [15]. The power curves were used straight from the manufacturers, and were not smoothed, like [8] for example recommends. The data removal was made only if it was obvious that there was some problem with the production not caused by a lack of wind. If these time spans would not have been removed, unnecessary bias would have been introduced into the results. After analysing the real production data, it seems plausible that production restrictions are imposed on the wind farms, since the production frequency for higher output seems far off of what it could or should be, as discussed in more detail in the closing chapter.

# 6. SUMMARY AND CONCLUSION

A simulation model for the wind power generation of 4 wind farms, 1 in Austria and 3 in New Zealand was developed by using MERRA reanalysis data and the power curves of the specific turbines installed in the wind farms. The MERRA data, gridded at a spatial resolution of 0.5° latitude and 0.66° longitude and resolved hourly, contains wind speeds for different heights. The wind velocities of the closest point of the MERRA dataset to each wind farm were extrapolated to the hub heights of the turbines. These wind speeds were used to feed the power curves and simulate the power generation of the wind farms. Afterwards the simulated power generation was compared to the real production data of the wind farms. The analysis examines hourly production as well as aggregated daily, monthly, quarterly and annual production. For the Austrian wind farm the total amount of production is underestimated and equals only 87.0% of the real production, whereby for White Hill it is overestimated by 14.3%. However, this overestimation could be a consequence of restrictions in production imposed to the operator, since the maximum production capacity is 58MW and the highest and second highest production (57.5MW, 56.2MW) were only reached once in 7 years, which is very unlikely if the wind farm is operated under normal, non-restricted conditions. The remaining 2 wind farms – Te Apiti and Mahinerangi – are covered quite well, concerning the total amount of production.

Nevertheless, the correlation coefficients (and their confidence intervals) show rather unsatisfying results for some wind farms and especially for higher temporal resolutions, i.e. hourly or daily. Even though for lower temporal resolutions, simulation and real production correlate quite well, the lack of a good correlation for hourly or daily resolution cannot be ignored. Still, not all wind farms perform equally. The best performance, regarding the correlation coefficients and the confidence intervals, is achieved by the Austrian wind farm. The 95% confidence interval for hourly production in Austria is 0.745 – 0.754 and for daily production 0.837 – 0.862, which represents the highest correlation values for hourly and daily resolutions. The worst results are achieved for Te Apiti, which only has a 95% confidence interval of 0.666 – 0.674 for hourly and 0.722 – 0.757 for daily production.

The results leave an ambivalent impression. On the one hand, the simulation model is not powerful enough to reflect the real wind power production in a satisfying way (regarding the total amount of production and the correlation values), in particular at high temporal resolution. MERRA reanalysis data is therefore not suitable for the assessment of individual locations in terms of profitability of wind turbine installations, as both the mean and the temporal profile of wind power production may be important, depending on the form of regulation. On the other hand, it can reproduce the monthly and seasonal production to a high extent, which is useful when the data is used in large scale integration studies. And besides that, the developed model provides a basis for further research, which could result in an optimized and improved model that could improve on quality of simulated wind power production.

# 7. REFERENCES

[1] B. Sørensen, *Renewable energy: physics, engineering, environmental impacts, economics & planning*, 4th ed. Burlington, MA: Academic Press, 2011.

[2] V. Quaschning, *Renewable energy and climate change*. Chichester, West Sussex, U.K. ; Hoboken, N.J: Wiley, 2010.

[3] Intergovernmental Panel on Climate Change, Hrsg., „Observations: Atmosphere and Surface“, in *Climate Change 2013 - The Physical Science Basis*, Cambridge: Cambridge University Press, 2014, S. 159–254.

[4] Intergovernmental Panel on Climate Change, *Climate Change 2014: Mitigation of Climate Change: Working Group III Contribution to the IPCC Fifth Assessment Report*. 2015.

[5] „Electricity production and supply statistics - Statistics Explained“. [Online]. Verfügbar unter: http://ec.europa.eu/eurostat/statistics-explained/index.php/Electricity_production_and_supply_statistics. [Zugegriffen: 27-Juni-2016].

[6] R. Lucchesi, „File Specification for MERRA Products. GMAO Office Note No. 1 (Version 2.3)“, 2012. [Online]. Verfügbar unter: http://gmao.gsfc.nasa.gov/products/documents/MERRA_File_Specification.pdf. [Zugegriffen: 20-Feb-2016].

[7] M. M. Rienecker, M. J. Suarez, R. Gelaro, R. Todling, J. Bacmeister, E. Liu, M. G. Bosilovich, S. D. Schubert, L. Takacs, G.-K. Kim, S. Bloom, J. Chen, D. Collins, A. Conaty, A. da Silva, W. Gu, J. Joiner, R. D. Koster, R. Lucchesi, A. Molod, T. Owens, S. Pawson, P. Pegion, C. R. Redder, R. Reichle, F. R. Robertson, A. G. Ruddick, M. Sienkiewicz, und J. Woollen, „MERRA: NASA’s Modern-Era Retrospective Analysis for Research and Applications“, *J. Clim.*, Bd. 24, Nr. 14, S. 3624–3648, Juli 2011.

[8] J. Olauson und M. Bergkvist, „Modelling the Swedish wind power production using MERRA reanalysis data“, *Renew. Energy*, Bd. 76, S. 717–725, Apr. 2015.

[9] D. J. Cannon, D. J. Brayshaw, J. Methven, P. J. Coker, und D. Lenaghan, „Using reanalysis data to quantify extreme wind power generation statistics: A 33 year case study in Great Britain“, *Renew. Energy*, Bd. 75, S. 767–778, März 2015.

[10] P. Juruš, K. Eben, J. Resler, P. Krč, I. Kasanický, E. Pelikán, M. Brabec, und J. Hošek, „Estimating climatological variability of solar energy production“, *Sol. Energy*, Bd. 98, Part C, S. 255–264, Dez. 2013.

[11] K. A. Janker, „Aufbau und Bewertung einer für die Energiemodellierung verwendbaren Datenbasis an Zeitreihen erneuerbarer Erzeugung und sonstiger Daten“, Technische Universität München, 2015.

[12] J. Schmidt, R. Cancella, und A. O. Pereira Jr., „An optimal mix of solar PV, wind and hydro power for a low-carbon electricity supply in Brazil“, *Renew. Energy*, Bd. 85, S. 137–147, Jan. 2016.

[13] M. Huber und C. Weissbart, „On the optimal mix of wind and solar generation in the future Chinese power system“, *Energy*, Bd. 90, Part 1, S. 235–243, Okt. 2015.

[14] M. B. McElroy, X. Lu, C. P. Nielsen, und Y. Wang, „Potential for Wind-Generated Electricity in China“, *Science*, Bd. 325, Nr. 5946, S. 1378–1380, Sep. 2009.

[15] I. Staffell und R. Green, „How does wind farm performance decline with age?“, *Renew. Energy*, Bd. 66, S. 775–786, Juni 2014.

[16] M. Ritter, Z. Shen, B. López Cabrera, M. Odening, und L. Deckert, „Designing an index for assessing wind energy potential“, *Renew. Energy*, Bd. 83, S. 416–424, Nov. 2015.

[17] „GeoHack - White Hill Wind Farm“. [Online]. Verfügbar unter: https://tools.wmflabs.org/geohack/geohack.php?pagename=White_Hill_Wind_Farm&params=45_45_9_S_168_16_18_E_type:landmark_region:NZ. [Zugegriffen: 22-Feb-2016].

[18] „http://www.windenergy.org.nz/white-hill-wind-farm“. [Online]. Verfügbar unter: http://www.windenergy.org.nz/white-hill-wind-farm. [Zugegriffen: 22-Feb-2016].

[19] „http://www.windenergy.org.nz/mahinerangi-wind-farm“. [Online]. Verfügbar unter: http://www.windenergy.org.nz/mahinerangi-wind-farm. [Zugegriffen: 22-Feb-2016].

[20] „Mahinerangi Wind Farm“, *Wikipedia, the free encyclopedia*. 22-März-2015.

[21] „GeoHack - Mahinerangi Wind Farm“. [Online]. Verfügbar unter: https://tools.wmflabs.org/geohack/geohack.php?pagename=Mahinerangi_Wind_Farm&params=45_45_38_S_169_54_18_E_type:landmark_region:NZ. [Zugegriffen: 22-Feb-2016].

[22] „Te Apiti Wind Farm“. [Online]. Verfügbar unter: http://www.windenergy.org.nz/te-apiti-wind-farm. [Zugegriffen: 24-Feb-2016].

[23] „GeoHack - Te Apiti Wind Farm“. [Online]. Verfügbar unter: https://tools.wmflabs.org/geohack/geohack.php?pagename=Te_Apiti_Wind_Farm&params=40_17_46_S_175_48_30_E_type:landmark. [Zugegriffen: 24-Feb-2016].

[24] „Te Apiti Wind Farm“, *Wikipedia, the free encyclopedia*. 06-Dez-2015.

[25] „Danish wind giants head for merger -- Vestas and NEG Micon going for global supremacy“. [Online]. Verfügbar unter: http://www.windpowermonthly.com/article/956160/danish-wind-giants-head-merger---vestas-neg-micon-going-global-supremacy. [Zugegriffen: 24-Feb-2016].

[26] H. Kraus, *Die Atmosphäre der Erde: Eine Einführung in die Meteorologie*. Springer-Verlag, 2007.

[27] „Enercon E70 2000kW Datasheet.pdf“. .

[28] „vestas v80.pdf“. [Online]. Verfügbar unter: http://www.kulak.com.pl/Wiatraki/pdf/vestas%20v80.pdf. [Zugegriffen: 06-Apr-2016].

[29] „NEG Micon NM 72/1650 - 1.650,0 kW - Turbine“. [Online]. Verfügbar unter: http://en.wind-turbine-models.com/turbines/1300-neg-micon-nm-72-1650. [Zugegriffen: 24-Feb-2016].

[30] I. Staffell, „Wind Turbine Power Curves“. [Online]. Verfügbar unter: https://www.academia.edu/1489838/Wind_Turbine_Power_Curves. [Zugegriffen: 20-Feb-2016].

[31] „ENERCON_Produkt_Innen_en_20032014.indd - ENERCON_Produkt_en_06_2015.pdf“. [Online]. Verfügbar unter: http://www.enercon.de/fileadmin/Redakteur/Medien-Portal/broschueren/pdf/en/ENERCON_Produkt_en_06_2015.pdf. [Zugegriffen: 06-Apr-2016].

[32] „Data Re-Processing — GES DISC - Goddard Earth Sciences Data and Information Services Center“. [Online]. Verfügbar unter: http://disc.sci.gsfc.nasa.gov/mdisc/documentation/data-re-processing. [Zugegriffen: 26-Feb-2016].

[33] J. Schallenberg-Rodriguez, „A methodological review to estimate techno-economical wind energy production“, *Renew. Sustain. Energy Rev.*, Bd. 21, S. 272–287, Mai 2013.

[34] N. S. Chok, „Pearson’s versus Spearman’s and Kendall’s correlation coefficients for continuous data“, University of Pittsburgh, 2010.

[35] D. G. Bonett und T. A. Wright, „Sample size requirements for estimating Pearson, Kendall and Spearman correlations“, *Psychometrika*, Bd. 65, Nr. 1, S. 23–28, 2000.

[36] „Confidence Interval for Pearson’s Correlation - Confidence_Interval_for_Pearsons_Correlation.pdf“. [Online]. Verfügbar unter: http://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/PASS/Confidence_Interval_for_Pearsons_Correlation.pdf. [Zugegriffen: 24-Mai-2016].

# 8. APPENDIX (R PROGRAM-CODES)

## 8.1. READING FUNCTIONS FOR PARAMETERS

```
library(ncdf4)
rad <- pi/180
########################## Reading Functions ###########################

datum <- function(ncname) {
  ncfile <- nc_open(ncname)
  h <- ncvar_get(ncfile, "time")
  d <- unlist(strsplit(ncfile$dim$time$units, " "))
  date <- rep(d[3],24)
  dh <- paste(date,h)
  x <- as.POSIXct(strptime(dh, format="%Y-%m-%d %H",tz="UTC"))
  nc_close(ncfile)
  return(x)
}

readu50m <- function(ncname) {
  du50m <- "u50m"
  ncfile <- nc_open(ncname)

  #Longitude
  longitude <- ncvar_get(ncfile, "longitude", verbose = F)
  nlon <- dim(longitude)

  #Latitude
  latitude <- ncvar_get(ncfile, "latitude", verbose = F)
  nlat <- dim(latitude)

  #Time
  time <- ncvar_get(ncfile, "time")
  tunits <- ncatt_get(ncfile, "time", "units")
  ntime <- dim(time)

  #read the variable
  u50m.array <- ncvar_get(ncfile, du50m)

  #Dataframe
  u50m.vec.long <- as.vector(u50m.array)
  u50m.mat <- matrix(u50m.vec.long, nrow = nlon * nlat, ncol = ntime)
  lonlat <- expand.grid(longitude,latitude)
  lonlat <- lonlat*rad

  # Distance between points
  dista <- 6378.388*acos(sin(lat) * sin(lonlat[,2]) + cos(lat) * 
cos(lonlat[,2]) * cos(lonlat[,1]-long))
  dfu50m <- (data.frame(cbind(dista,lonlat/rad, u50m.mat)))
  dfu50m <- dfu50m[ order(dfu50m[,1]), ]
  names(dfu50m) <- c("dista", "Longitude", "Latitude", seq(1:24))
  nc_close(ncfile)
  return(dfu50m)
}

readv50m <- function(ncname) {
  dv50m <- "v50m"
  ncfile <- nc_open(ncname)

  #Longitude
  longitude <- ncvar_get(ncfile, "longitude", verbose = F)
  nlon <- dim(longitude)

  #Latitude
  latitude <- ncvar_get(ncfile, "latitude", verbose = F)
  nlat <- dim(latitude)

  #Time
  time <- ncvar_get(ncfile, "time")
  tunits <- ncatt_get(ncfile, "time", "units")
  ntime <- dim(time)

  #read the variable
  v50m.array <- ncvar_get(ncfile, dv50m)

  #Dataframe
  v50m.vec.long <- as.vector(v50m.array)
  v50m.mat <- matrix(v50m.vec.long, nrow = nlon * nlat, ncol = ntime)
  lonlat <- expand.grid(longitude,latitude)
  lonlat <- lonlat*rad

  # Distance between points
  dista <- 6378.388*acos(sin(lat) * sin(lonlat[,2]) + cos(lat) *
cos(lonlat[,2]) * cos(lonlat[,1]-long))
  dfv50m <- (data.frame(cbind(dista,lonlat/rad, v50m.mat)))
  dfv50m <- dfv50m[ order(dfv50m[,1]), ]
  names(dfv50m) <- c("dista", "Longitude", "Latitude", seq(1:24))
  nc_close(ncfile)
  return(dfv50m)
}

readu10m <- function(ncname) {
  du10m <- "u10m"
  ncfile <- nc_open(ncname)

  #Longitude
  longitude <- ncvar_get(ncfile, "longitude", verbose = F)
  nlon <- dim(longitude)

  #Latitude
  latitude <- ncvar_get(ncfile, "latitude", verbose = F)
  nlat <- dim(latitude)

  #Time
  time <- ncvar_get(ncfile, "time")
  tunits <- ncatt_get(ncfile, "time", "units")
  ntime <- dim(time)

  #read the variable
  u10m.array <- ncvar_get(ncfile, du10m)

  #Dataframe
  u10m.vec.long <- as.vector(u10m.array)
  u10m.mat <- matrix(u10m.vec.long, nrow = nlon * nlat, ncol = ntime)
  lonlat <- expand.grid(longitude,latitude)
  lonlat <- lonlat*rad

  # Distance between points
  dista <- 6378.388*acos(sin(lat) * sin(lonlat[,2]) + cos(lat) *
cos(lonlat[,2]) * cos(lonlat[,1]-long))
  dfu10m <- (data.frame(cbind(dista, lonlat/rad, u10m.mat)))
  dfu10m <- dfu10m[ order(dfu10m[,1]), ]
  names(dfu10m) <- c("dista", "Longitude", "Latitude", seq(1:24))
  nc_close(ncfile)
  return(dfu10m)
}

readv10m <- function(ncname) {
  dv10m <- "v10m"
  ncfile <- nc_open(ncname)

  #Longitude
  longitude <- ncvar_get(ncfile, "longitude", verbose = F)
  nlon <- dim(longitude)

  #Latitude
  latitude <- ncvar_get(ncfile, "latitude", verbose = F)
  nlat <- dim(latitude)

  #Time
  time <- ncvar_get(ncfile, "time")
  tunits <- ncatt_get(ncfile, "time", "units")
  ntime <- dim(time)

  #read the variable
  v10m.array <- ncvar_get(ncfile, dv10m)

  #Dataframe
  v10m.vec.long <- as.vector(v10m.array)
  v10m.mat <- matrix(v10m.vec.long, nrow = nlon * nlat, ncol = ntime)
  lonlat <- expand.grid(longitude,latitude)
  lonlat <- lonlat*rad

  # Distance between points
  dista <- 6378.388*acos(sin(lat) * sin(lonlat[,2]) + cos(lat) *
cos(lonlat[,2]) * cos(lonlat[,1]-long))
  dfv10m <- (data.frame(cbind(dista, lonlat/rad, v10m.mat)))
  dfv10m <- dfv10m[ order(dfv10m[,1]), ]
  names(dfv10m) <- c("dista", "Longitude", "Latitude", seq(1:24))
  nc_close(ncfile)
  return(dfv10m)
}

readu2m <- function(ncname) {
  du2m <- "u2m"
  ncfile <- nc_open(ncname)

  #Longitude
  longitude <- ncvar_get(ncfile, "longitude", verbose = F)
  nlon <- dim(longitude)

  #Latitude
  latitude <- ncvar_get(ncfile, "latitude", verbose = F)
  nlat <- dim(latitude)

  #Time
  time <- ncvar_get(ncfile, "time")
  tunits <- ncatt_get(ncfile, "time", "units")
  ntime <- dim(time)

  #read the variable
  u2m.array <- ncvar_get(ncfile, du2m)

  #Dataframe
  u2m.vec.long <- as.vector(u2m.array)
  u2m.mat <- matrix(u2m.vec.long, nrow = nlon * nlat, ncol = ntime)
  lonlat <- expand.grid(longitude,latitude)
  lonlat <- lonlat*rad

  # Distance between points
  dista <- 6378.388*acos(sin(lat) * sin(lonlat[,2]) + cos(lat) *
cos(lonlat[,2]) * cos(lonlat[,1]-long))
  dfu2m <- (data.frame(cbind(dista, lonlat/rad, u2m.mat)))
  dfu2m <- dfu2m[ order(dfu2m[,1]), ]
  names(dfu2m) <- c("dista", "Longitude", "Latitude", seq(1:24))
  nc_close(ncfile)
  return(dfu2m)
}

readv2m <- function(ncname) {
  dv2m <- "v2m"
  ncfile <- nc_open(ncname)

  #Longitude
  longitude <- ncvar_get(ncfile, "longitude", verbose = F)
  nlon <- dim(longitude)

  #Latitude
  latitude <- ncvar_get(ncfile, "latitude", verbose = F)
  nlat <- dim(latitude)

  #Time
  time <- ncvar_get(ncfile, "time")
  tunits <- ncatt_get(ncfile, "time", "units")
  ntime <- dim(time)

  #read the variable
  v2m.array <- ncvar_get(ncfile, dv2m)

  #Dataframe
  v2m.vec.long <- as.vector(v2m.array)
  v2m.mat <- matrix(v2m.vec.long, nrow = nlon * nlat, ncol = ntime)
  lonlat <- expand.grid(longitude,latitude)
  lonlat <- lonlat*rad

  # Distance between points
  dista <- 6378.388*acos(sin(lat) * sin(lonlat[,2]) + cos(lat) *
cos(lonlat[,2]) * cos(lonlat[,1]-long))
  dfv2m <- (data.frame(cbind(dista, lonlat/rad, v2m.mat)))
  dfv2m <- dfv2m[ order(dfv2m[,1]), ]
  names(dfv2m) <- c("dista", "Longitude", "Latitude", seq(1:24))
  nc_close(ncfile)
  return(dfv2m)
}

readdisph <- function(ncname) {
  disph <- "disph"
  ncfile <- nc_open(ncname)

  #Longitude
  longitude <- ncvar_get(ncfile, "longitude", verbose = F)
  nlon <- dim(longitude)
  #Latitude
  latitude <- ncvar_get(ncfile, "latitude", verbose = F)
  nlat <- dim(latitude)

  #Time
  time <- ncvar_get(ncfile, "time")
  tunits <- ncatt_get(ncfile, "time", "units")
  ntime <- dim(time)

  #read the variable
  disph.array <- ncvar_get(ncfile, "disph")

  #Dataframe
  disph.vec.long <- as.vector(disph.array)
  disph.mat <- matrix(disph.vec.long, nrow = nlon * nlat, ncol = ntime)
  lonlat <<- expand.grid(longitude,latitude)
  lonlat <- lonlat*rad

  # Distance between points
  dista <- 6378.388*acos(sin(lat) * sin(lonlat[,2]) + cos(lat) *
cos(lonlat[,2]) * cos(lonlat[,1]-long))
  dfdisph <- (data.frame(cbind(dista, lonlat/rad, disph.mat)))
  dfdisph <- dfdisph[ order(dfdisph[,1]), ]
  names(dfdisph) <- c("dista", "Longitude", "Latitude", seq(1:24))
  nc_close(ncfile)
  return(dfdisph)
}
```

## 8.2. SIMULATION OF WHITE HILL

```
library(ncdf4)

setwd ("C:/Users/mesa-/OneDrive/Master/Masterarbeit/MERRA/MERRA Wind
Neuseeland")


######################### Windpark White Hill #########################

#----------------------------------------------------------------------#

############# 29 V90 Turbinen seit Beginn (01.06.2007) in Betrieb
##########


# https://www.meridianenergy.co.nz/about-us/our-power-stations/wind/white-
hill

# http://www.windenergy.org.nz/white-hill-wind-farm

# Operator of White Hill


# Degree East / North

#
https://tools.wmflabs.org/geohack/geohack.php?pagename=White_Hill_Wind_Farm
&params=45_45_9_S_168_16_18_E_type:landmark_region:NZ


long <- 168.271667*rad

lat <- -45.7525*rad


# Lists for variables and all files


NZfiles <- list.files(pattern = "*.nc")

Listdate <- lapply(NZfiles, datum)

Listu50mWH <- lapply(NZfiles, readu50m)

Listv50mWH <- lapply(NZfiles, readv50m)

Listu10mWH <- lapply(NZfiles, readu10m)

Listv10mWH <- lapply(NZfiles, readv10m)

Listu2mWH <- lapply(NZfiles, readu2m)

Listv2mWH <- lapply(NZfiles, readv2m)

ListdisphWH <- lapply(NZfiles, readdisph)


# Nearest Neighbor Interpolation Matrix for all files


NNdate <- unlist(Listdate)

NNWHu50m <- unlist(sapply(Listu50mWH, function(d) d[1,4:27]))

NNWHv50m <- unlist(sapply(Listv50mWH, function(d) d[1,4:27]))
```

```
NNWHu10m <- unlist(sapply(Listu10mWH, function(d) d[1,4:27]))

NNWHv10m <- unlist(sapply(Listv10mWH, function(d) d[1,4:27]))

NNWHu2m <- unlist(sapply(Listu2mWH, function(d) d[1,4:27]))

NNWHv2m <- unlist(sapply(Listv2mWH, function(d) d[1,4:27]))

NNWHdisph <- unlist(sapply(ListdisphWH, function(d) d[1,4:27]))



#####################Empty Workspace###############

rm(Listu2mWH,Listv2mWH,Listu10mWH,Listv10mWH,Listv50mWH,Listu50mWH,Listdisp
hWH)



# MERRA date

MD <- as.POSIXct(NNdate, origin = "1970-01-01 00:00:00 UTC",
tz="Etc/Universal")


# Wind speeds Nearest Neighbor


WHuv50 <- sqrt(NNWHu50m^2+NNWHv50m^2)

WHuv10 <- sqrt(NNWHu10m^2+NNWHv10m^2)

WHuv2 <- sqrt(NNWHu2m^2+NNWHv2m^2)


#### WH MERRA-DF ###

MWH <- data.frame(MD,WHuv50,WHuv10,WHuv2,NNWHdisph)


# NZ Windfarm - data hourly and in MW!


setwd("C:/Users/mesa-/OneDrive/Master/Masterarbeit")

NZ.csv <- read.csv("DataNZAG_windpower.csv")


NZdate <- as.POSIXct(strptime(NZ.csv[,1], format="%Y-%m-%d
%H:%M:%S",tz="NZ"))

WH <- na.omit(data.frame(NZ.csv[,6]*1000,NZdate))

WH[,2] <- as.POSIXct(WH[,2], origin = "1970-01-01 00:00:00 UTC",
tz="Pacific/Auckland")

names(WH) <- c("WH Production","NZdate")


# Merge Data-Frames and drop out data

# first 3000 > Starting Process , 31000 - 32750 maintenance > drop out


WHDF <- merge(WH, MWH, by.x = "NZdate", by.y = "MD")

length(WHDF[,1])
```

```
WHDF <- WHDF[c(3001:30999,32751:51122),]


### check for timezone/merge

utils::View(MWH)

head(WHDF)

# check first row of head(WHDF) > 4.10.07 17:00 uv50=10.07 and same date of
MWH >>> different values > 13 hours earlier > same value



######################  Nearest Neighbor Modelling
##########################


# Power Curve Vestas V80

# turbine model   Vestas: V80-2.0MW
https://en.wikipedia.org/wiki/White_Hill_Wind_Farm

# Data  http://www.kulak.com.pl/Wiatraki/pdf/vestas%20v80.pdf

# Paper Staffel Power Curves

V80power <-
c(0,0,0,66.3,152,280,457,690,978,1296,1598,1818,1935,1980,1995,1999,2000,20
00,2000,2000,2000,2000,2000,2000,2000)

V80wind <-
c(0,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25)


# alpha / friction coefficient

a1050WH <- (log(WHDF[,3])-log(WHDF[,4]))/(log(50)-log(10+WHDF[,6]))


# Wind speeds with power law from 50 to 67 extrapolated

v6750WH <- WHDF[,3]*(67/50)^a1050WH

matplot(data.frame(v6750WH,WHDF[,3]),type="l")



# Function for power curve

fWH <- approxfun(V80wind, V80power)

WHcurve <- curve(fWH(x),0, 25, col = "green2", lwd=4,main="Vestas V80-
2.0MW", xlab="Wind speed m/s", ylab="Power in kW", n=25,font=4)


# Production

WHP <- sum(WHDF[,2])              # Sum White Hill real produktion (29
Turbines)

WHpl1 <- sapply(v6750WH, FUN = "fWH")    # modelled production with power
law (alpha) for 1 turbine

WHpl <- 29*WHpl1          # hourly modelled Production 29 Turbines

WHplsum <- sum(WHpl) # sum of modelled production
```

```
WHplsum/WHP # real:modelled production (sum)

#Correlation hourly
cor(WHDF[,2],WHpl)
```

## 8.3. ANALYSIS OF WHITE HILL

```
########### White Hill ############

# Dataframe with date, real and modelled production
WHA <-cbind(WHDF[,1:2],WHpl)
WHA[,1] <- as.POSIXct(WHDF[,1],origin = "1970-01-01 10:00:00")
names(WHA) <- c("Datum","Echt","Modell")
head(WHA)
head(WHDF)

# List / Character White Hill
WHY <- format(WHA[,1],"%Y")
WHYm<-format(WHA[,1],"%Y%m")
WHYmd<-format(WHA[,1],"%Y%m%d")
WHm<-format(WHA[,1],"%m")
WHd<-format(WHA[,1],"%d")
WHq<-quarter(WHA[,1],with_year = TRUE)
utils::View(WHm)

# Aggregte White Hill
ag_yearWH<- aggregate(WHA[,2:3],by=list(WHY),sum)
ag_monWH<-aggregate(WHA[,2:3],by=list(WHYm),sum)
ag_dayWH<-aggregate(WHA[,2:3],by=list(WHYmd),sum)
ag_seasWH<-aggregate(WHA[,2:3],by=list(WHm),sum)
ag_dWH<-aggregate(WHA[,2:3],by=list(WHd),sum)
ag_qWH<-aggregate(WHA[,2:3],by=list(WHq),sum)

#length
length(WHDF[,2]) #46371 Stunden
length(ag_dayWH[,2]) #1934 tage
length(ag_monWH[,2]) #65 Monate
length(ag_qWH[,2]) #22 Quartale
length(ag_yearWH[,2]) #7Jahre

########## Graphics / correlations White Hill  #############
# years
matplot(ag_yearWH[,2:3],type="l")
cor(ag_yearWH[,2:3])
cor(ag_yearWH[,2:3],method="spearman")

# single months
matplot(ag_monWH[,2:3],type="l")
matplot(ag_monWH[,2:3]/1000,type="l",ylab="Production in MWh",
        lwd = 3,col = c("lightslateblue","red"),lty = 1,main="Monthly
Production White Hill",
        xlab="Months",cex.axis=1.25,cex.lab=1.25)
legend("topright",c("Simulated","Real"),col=c("Red","lightslateblue"),lwd=1
0)
legend("top",paste("r","=","0.8860213"))
cor(ag_monWH[,2:3])

# single days
matplot(ag_dayWH[,2:3],type="l")
cor(ag_dayWH[,2:3])

# aggregated months
matplot(ag_seasWH[,2:3],type="l")
cor(ag_seasWH[,2:3])

# aggregated calendar days (1.-31.)
matplot(ag_dWH[,2:3],type="l")
cor(ag_dWH[,2:3])

# aggregated quarters
matplot(ag_qWH[,2:3],type="l")
cor(ag_qWH[,2:3])
cor(ag_qWH[,2:3],method="spearman")

#hours
cor(WHA[,2:3])

#### Modal Value, Skewness, RMSE, SD
library(modeest)
mlv(WHA[,2],method="naive")
mlv(WHA[,3],method="naive")
skewness(WHA[,2])

# White Hill capacity 58000kW
WHR<- WHA[,2]/58000
```

```
WHS<- WHA[,3]/58000
summary(WHR)-summary(WHS)
sd(WHR)
sd(WHS)
sd(WHR)-sd(WHS)
summary(WHR)
summary(WHS)
CIr(0.7,46371,level=0.95)  # hourly
CIr(0.8,1934,level=0.95)  # daily
CIr(0.87,65,level=0.95) # monthly
CIr(0.86,22,level=0.95) # seasonally
CIr(0.98,7,level=0.95) # annually
################ White Hill

# distance and u50m value table
setwd ("C:/Users/mesa-/OneDrive/Master/Masterarbeit/MERRA/MERRA Wind
Neuseeland")
nc <- "MERRA300.prod.assim.tavg1_2d_slv_Nx.20050101.SUB.nc"
long <- 168.271667*rad
lat <- -45.7525*rad
WHdist <- readu50m(nc)
WHdist[1:5,1:6]

# production in hours // assumption start-up time + maintainance times
plot(WH[,1],type="l", col="cadetblue4", lty=3,main= "Electricity production
White Hill Wind Farm from 01.06.2007 to 31.03.2013", xlab = "Hours",
ylab="Production in kWh")
####### first 3000h start-up , maintainance from 31000 to 32750
plot(NNDFWH[33000:33750,6],type="l")

# Maxima
which.max(WH[,1])
WH[18975,1]
WH[,1][WH[,1]>56000]

# histogram


par(mfrow=c(1,1))
WHhist <- hist(WHA[,3], breaks=seq(0,60000,by=3000),main="White Hill
Production",xlab="kWh",col = "red")
WHplhist <- hist(WHA[,2],breaks=seq(0,60000,by=3000),main="White Hill
Simulated Production",xlab="",col=mycol,add=T)
```

```
legend("topright",c("Simulated","Real"),WHhist$counts-
WHplhist$counts,col="green", lwd=10)

legend("topright",c("Simulated", "Real"),col=c("red",mycol),lwd=10)

sum(WHA[,2]>56000)


mycol <- rgb(0, 100, 255, max = 255, alpha = 100, names = "blue50")


# histogram 2d contour  package plot_ly

e1<-plot_ly(x=WHA[,2],type="histogram")

e2<-plot_ly(x=WHA[,3],type="histogram")

e3<-
plot_ly(x=WHA[,2],y=WHA[,3],type="histogram2dcontour",autocontour=FALSE,

            contours = list(coloring="fill",start=0,end=800,size=50))

e3

e1

e2

layout(e3,

       title="2dcontourplot // Bins = 20 (0-400)",

       yaxis=list(title="Simulated Production", rangemode="nonnegative"),

       xaxis=list(title="Real Production", rangemode="nonnegative")

)


xaxis = list(rangemode = "tozero"),

yaxis = list(rangemode = "nonnegative"))
```