UPTEC W10 003
Examensarbete 30 hp
Mars 2010

Statistical analysis of wind data regarding long-term correction

Statistisk analys av vinddata med avseende på långtidskorrigering

Christoffer Jonsson

ABSTRACT

STATISTICAL ANALYSIS OF WIND DATA REGARDING LONG-TERM CORRECTION

Christoffer Jonsson

The procedure of determining if a site is suitable for wind power production requires convincing statistical data describing the long-term behavior of the average wind speed. This can be achieved by measuring the wind speed for a short time period, e.g. a year, and after that a Measure-Correlate-Predict (MCP) method can be performed. The short-term measured wind data must be used in combination with a long-term reference series. This long-term reference series can be global reanalysis data reaching 20 to 30 years back in time. In a MCP method different regression methods can be used. After creating a long-term corrected wind data series, it is possible to analyze the conditions at the investigated site. To be able to study the behavior of different reference series and regression methods, a model was created in MATLAB. As short-term wind speed data Vattenfall Wind Power supplied data from two measuring masts, Ringhals and Oskarshamn, with maximum heights of 96 and 100 meters, respectively. From Uppsala University data were supplied from a measuring mast near Marsta with maximum measurement height of 29 meters.

When creating these long-term corrected wind data series there were many methods available. In this Master thesis methods such as Ordinary-Least-Square, Least-Absolute-Deviation and Reduced-Major-Axis regression methods have been used. With each method three reference series were used in combination with the short-term measurement data. These were data from NCAR 850 hPa, NCAR 42-meter sigma level and a confidential source.

Regression methods in combination with reference series were studied and the deviation from mean wind speed was obtained for each of these cases. Studies were performed on how the length of the short-term measurement series affected the deviation from the measured mean wind speed. It was also investigated if the time of the year had any influence on the measurements.

The general conclusion drawn after performing the above-mentioned studies was that the NCAR 850 hPa wind speed data and the Reduced-Major-Axis regression method gave the smallest deviation from the measured mean wind speed in most cases. It was also concluded that when a short-term measurement series reached 10 to 14 months there was a significant decrease in deviation from the mean wind speed, regardless of reference series or method used. Calculations from the model regarding seasonal dependence stated that there was a slight dependency on which period of the year a measurement was performed.

Keywords: MATLAB, Measure-Correlate-Predict, Regression, NCAR/NCEP, Seasonal dependence, Long-term Correction, Wind Analysis

Department of Earth Sciences, Geocentrum, Villavägen 16, SE-752 36 UPPSALA

ISSN 1401-5765

REFERAT

STATISTISK ANALYS AV VINDDATA MED AVSEENDE PÅ LÅNGTIDSKORRIGERING

Christoffer Jonsson

I processen att bedöma om en plats är lämplig för utbyggnad av vindkraft måste det finnas övertygande statistiska data som beskriver den genomsnittliga vindhastigheten över en längre tid. Genom att utföra vindhastighetsmätningar på den tänkta platsen under en kortare tid, exempelvis ett år, och därefter tillämpas en Measure-Correlate-Predict (MCP) metod i kombination med en långtidsreferens, exempelvis en global modell som sträcker sig 20 till 30 år bakåt i tiden kan detta göras. I en MCP-metod kan olika typer av regressionsmetoder användas. När en långtidskorrigerad vinddataserie finns tillgänglig kan dess beteende på den tänkta platsen analyseras. För att kunna göra detta för flera olika typer av referensserier och regressionsmetoder skapades en modell i MATLAB. Två vinddataserier erhölls från Vattenfall Vindkraft. Dessa var Ringhals och Oskarshamn med högsta mäthöjd på 96 respektive 100 meter. En ytterligare vinddataserie erhölls av Uppsala Universitet från en mätmast nära Marsta med högsta mäthöjd på 29 meter.

Det fanns flera metoder tillgängliga för att skapa de långtidskorrigerade vinddataserierna. I det här examensarbetet har metoderna Ordinary-Least-Square-, Least-Absolute-Deviation- och Reduced-Major-Axis regressioner använts. För varje metod testades tre referensserier i kombination med de kortare vinddataserierna. Dessa var NCAR 850 hPa vindhastigheter, NCAR 42 meters sigmanivå vindhastigheter och annan meteorologisk data.

Regressionsmetoderna utvärderades genom att avvikelsen från de kortare mätseriernas medelvindhastigheter beräknades. Det undersöktes också hur längden på använd vinddata från de kortare mätserierna påverkade avvikelsen i medelvindhastighet och om det fanns något säsongsberoende på när under året som mätningen av vinddata var gjord.

Slutsatserna från undersökningarna var att NCAR 850 hPa vindhastigheter och regressionsmetoden Reduced-Major-Axis generellt gav de lägsta avvikelserna från uppmätt medelvindhastighet. Slutsatser kunde också dras om längden av använd mätdata. Det var tydligt att oavsett referensserie och regressionsmetod uppstod en minskningen i avvikelse från medelvindhastigheten mellan 10 till 14 månaders längd på mätserien. Resultat angående säsongsberoende kunde påvisas i form av avvikelser mellan mätningar gjorda under olika tidpunkter på året. Storlek och tecken på avvikelsen berodde på vilken referensserien i kombination med regressionsmetod som användes.

Nyckelord: MATLAB, Measure-Correlate-Predict, Regression, NCAR/NCEP, Säsongsberoende, Långtidskorrigering, Vindanalys

Institutionen för geovetenskaper, Geocentrum, Villavägen 16, SE-752 36 UPPSALA

ISSN 1401-5765

PREFACE

This Master thesis performed at Vattenfall Wind Power has given me the opportunity to deepen my knowledge in the field of wind energy assessment. Thanks to the prior coursework in empirical modeling and wind power energy, this Master thesis has become rather concrete and has put focus on statistical analysis of wind data. Much thanks to the people working at the technology department of Vattenfall Wind Power, Daniel Gustafsson, Jan-Åke Dahlberg, Måns Håkansson, Sven-Erik Thor and Staffan Snis.

Many thanks also to Hans Bergström who has been subject reviewer on behalf of Uppsala University, Department of Earth Sciences and Daniel Gustafsson supervising the work at Vattenfall Wind Power.

Stockholm, January 2010

Christoffer Jonsson

Printed at the Department of Earth Sciences, Geotryckeriet, Uppsala University, Uppsala 2010

POPULÄRVETENSKAPLIG SAMMANFATTNING

Statistisk analys av vinddata med avseende på långtidskorrigering

För att kunna göra en uppskattning av hur stor ekonomisk vinst som är möjlig att erhålla från ett planerat vindkraftsprojekt, oavsätt om det gäller ett fåtal vindkraftverk eller en hel vindkraftspark, är det oerhört viktigt att veta hur mycket det kommer att blåsa på den tänkta platsen. Med modellen som har utvecklats inom det här examensarbetet har osäkerheten i uppskattningarna av vindhastigheter i vissa fall kunnat minskas avsevärt. Detta har gjorts genom att använda olika regressionsmetoder för att beskriva förhållandet mellan uppmätt vindhastighet från den tänkta platsen och vindhastigheter från globala vädermodeller.

Genom att använda en modell som använder flera olika metoder i kombination med flera olika vindhastighetsserier kan man enkelt och tydligt se vilken metod i kombination med vilken serie som ger minst osäkerheter i en vindanalys. Betydelsen av en riktig vindanalys kan bättre förstås om man tänker på att ett vindkraftverk ska stå på samma plats i 15 – 20 år. Det modellen ytterligare påvisar är att det verkar finnas samband mellan när på året som mätningarna av vindhastigheten är gjorda och hur bra vindanalysen blir. Det innebär ytterligare ett moment att verdera i sina vindanalys.

Utgångspunkten i början av examensarbetet var att försöka få svar på ett antal frågor gällande vindanalys. Den första frågan som diskuterades inom ramen för examensarbetet var om trycknivåerna i NCEP/NCAR väderdatan skulle kunna användas som en ytterligare resurs när det gäller att uppskatta hur mycket det blåser på en plats och hur bra den i såfall skulle vara. Den andra frågan rörde hur länge man måste mäta vindhastigheten på en plats för att kunna vara säker på hur det blåser. Det fanns tidigare studier som antyder att efter 10 – 12 månader avtar förbättringsgraden avseevärt vad det gäller att kunna förutspå hur det blåser på en plats. Den tredje frågan var om och isåfall hur stor påverkan säsongsberoendet har på mätningarna och resultaten. Det vill säga, skulle det spelar någon roll om man gör en mätning i Januari eller Februari.

Det här examensarbetet resulterade i svar på dessa ovan nämnda frågor. Det visade sig att väderdata från trycknivåerna i NCEP/NCAR var relevanta resurser i vindanalysen. Det kunde också ses att avvikelsen från det mätta värdet minskade med tiden. Dock skedde en utplaning runt 10 till 14 månader, beroende på väderdata och metod. Detta innebär att om den mest ekonomiska vindanaylen ska kunna utföras måste 10 till 14 månaders mätdata erhållas från platsen. Efter utplaning minskar dock avvikelsen ytterligare men med en bestämt mindre takt. Testerna för att påvisa säsongsberoende visade att det fanns ett beroende men inga fler slutsatser kunde dras, detta då endast tre mätmaster användts. Det verkar även här spela roll vilken väderdata och vilken metod som används.

Avvikeser i vindanalysen beror alltså på använd metod, använd global vädermodell och på när själva mätningen av vindhastigheten är gjord på platsen. Detta examensarbete har påvisat komplexiteten i en vindanalys och att det finns ytterligare faktorer att studera. Här har endast den statistiska analysen gjorts. Det har alltså inte tagits hänsyn till hur de

globala modellerna gör sina beräkningar eller till hur resultaten kan förklaras meteorologiskt.

ACRONYMS

Acronym	Term
CDAS	Climate Data Assimilation System
hPa	hekto Pascal
LAD	Least-Absolute-Deviation
MCP	Measure-Correlate-Predict
NCAR	National Centre for Atmospheric Research
NCEP	National Centre for Environmental Prediction
OLS	Ordinary-Least-Square
RMA	Reduced-Major-Axis

CONTENTS

1 INTRODUCTION

The amount of wind power integrated in the Swedish energy production is growing rapidly. The development is driven by the economic interests of the major energy producers and additional goals set by the government. To be successful in planning and building wind farms, knowledge about how the wind behaves on-site is crucial. Knowing how much energy the wind contains on a planned site makes it possible to decide whether a project will succeed economically or not.

It is of importance to minimize all uncertainties in a wind assessment. The uncertainty in the long-term mean wind speed is the most important one to have a full understanding of. It is the long-term mean wind speed that determines how much electricity a wind turbine can produce. There are additional parameters that need to be investigated and evaluated, but without a high enough wind speed it will not be economically sustainable to build a wind turbine on the site.

Commonly, a Measure-Correlate-Predict method is used to determine the long-term behavior on-site. To make this procedure as effective as possible, there needs to be measurements on-site combined with a reference series. Treating these wind data with regression methods gives an understanding of the long-term behavior. From such a calculation, a wide range of parameter values can be evaluated. The Measure-Correlate-Predict method is accepted and proven to work well in this context.

No studies have been found in which one has investigated the effect of changing the regression method within the Measure-Correlate-Predict method and what effect that has on the uncertainty. Regarding reference series there are studies made on a specific reference series (AWS Truewind, 2006), but not testing different reference series against each other, combined with alternating the regression method. There are two types of reference series, one based on pressure levels and another on sigma levels. The influence on the uncertainty when using these two types of reference data has been tested separately. Studies where the reference series are changed and the results are compared to each other are rare. Pressure level data has been tested as reference series in wind assessments by both Nilsson & Bergström (2009) and AWS Truewind (2006).

The aim of this thesis was to clarify the effect of making changes in the Measure-Correlate-Predict method and describe these effects with hands-on statistical variables. The changes were regarding the regression method and the reference series. Three main questions were asked in the thesis work by using data from three different sites:

- Which effect has a change of the regression method in the Measure-Correlate-Predict method?
- Are there statistical parameters to describe the differences between reference series?
- Is the result of the Measure-Correlate-Predict method depending on during which season the measurements were made?

1.1 SCOPE OF THIS MASTER THESIS

To be able to validate how significant a method is for determining the long-term corrected wind speed, numerous parameters can be verified. The main goal of this Master thesis is to study different wind measurement series with different curve fittings.

In combination with preprocesses there might be improvements in the methodology used to calculate the long-term correction of the wind data. Different sources of wind data are tested against each other, e.g. measured wind speeds from measuring masts against reference series from virtually simulated global wind speeds for different heights and pressure levels. The focus is on regression and distribution methods for testing the significance of the reference series.

1.2 DEMARCATION

After the first theoretical phase numerous methods and usable parameters had been found. Only regression and resampling methods have been used in calculations. No considerations have been taken concerning the direction, temperature or other parameters included in many of the data sets.

2 BACKGROUND

2.1 A HISTORICAL REVIEW OF WIND POWER DEVELOPMENT

2.1.1 Emergence of the wind power industry

During the latter part of the 20th century, the climate change was closely observed and a more efficient usage of energy was indicated as one of the main areas for improvement to create a more sustainable environment. One of the most powerful European initiatives to create an environment containing a long-term sustainable production of electricity is the massive wind power development. The Swedish Government and the Swedish Energy Agency have declared high planning goals in comparison to the initial conditions of the Swedish wind power market. By the year 2020 Sweden should have at least an annual production of 20 TWh from wind power onshore and 10 TWh produced offshore (Energimyndigheten, 2009). Similar goals are stated in most of the European countries, which has resulted in an exponential growth of wind power plants being built during the last ten years (EWEA, 2008).

2.1.2 Development areas

One of the initial steps in planning and building a wind power plant or a wind farm is to consider the wind potential on site. The wind potential is the theoretical amount of energy that the wind contains, measured in the unit watts per square meter (W/m^2) (Wizelius, 2007). There is a strong need for the wind power industry to know how much wind energy a site contains. Today there are relatively good measurement and correlation methods for determining the wind energy content. One question that not yet has a clear answer is how long a measuring series has to be kept running to statistically represent the wind environment at the measuring site. The methods for determining and deducing the outcome of statistically treated data need to be more accurate and more clearly defined. Also there is a need for a way to quantify the quality of the reference data used in wind estimation modeling.

2.1.3 Further development

It is of great importance to understand and be able to determine the quality of the outcome from applied statistical methods. This is the core in further development of wind energy estimation methods. There might also be an unexplored assessment in frequency and spectral analyses of wind data series. Methods based on frequency instead of time detect smaller variations and trends in a much clearer way than regression models (Siddiqi et al., 2005).

2.2 WIND DATA COLLECTION

2.2.1 National Centers for Environmental Prediction and National Centre for Atmospheric Research

Two American organizations can be credited for the first attempts to build a global model for generation of weather data. In 1989 the National Centers for Environmental Prediction (NCEP) wanted to build a Climate Data Assimilation System (CDAS) based on a solid system with observations as input variables. The idea was to collect and store data so it would be accessible and editable in different computer programs for different scientific purposes. Observations from different types of measuring sources e.g. airplanes, wind masts and satellite observations were gathered and inserted in the model (Kistler et al., 1999). A dialog started with the National Centre for Atmospheric Research (NCAR) in 1990 and an agreement of starting a long reanalysis project, named "Reanalysis-1", was reached. In a reanalysis project collected data from the past, in combination with more recent data and actual results, are analyzed and composed into a new and enhanced data set. During four years, between 1990 and 1994, the development of the system occurred and 40 years of reanalyzed data were composed. One month of data was reanalyzed for each day the model ran, and the model was kept running for four years. During the next five years different parts of complementary data from 1948 to the present and continuing, were reanalyzed and corrected from useless data. This enabled a data series of over 50 years (Kistler et al., 1999).

In 1979 data from global satellites were introduced, which gave the NCEP/NCAR output data greater accuracy. There were still errors in the Reanalysis-1 data, some inevitable, like changes in the measurement equipment and improvement of the model and its parameters (Kistler et al., 1999). In the early stage of the NCAR/NCEP Reanalysis-1 project the quality of the collected data had a wide dispersion. There was therefore a second reanalysis of the data called the NCEP/DOE AMIP-II Reanalysis, or more commonly Reanalysis-2 (NOAA, 2005). The Reanalysis-2 began when satellites were added to the model, in the year 1979, and has been continuing non-stop, which gives the Reanalysis-2 data a time series of more than 30 years. The effect of adding satellite surveillance gave the model a sharper and higher resolution due to cloud detection and better heat surveillance (Kalnay et al., 1996).

2.2.2 Other Meteorological Data

Some content was excluded due to confidential material. In the official version this content will be called Other Data or Other Meteorological Data.

2.3 INTRODUCTION TO WIND MEASURMENT

There are a lot of meteorological institutes and other agencies that have been measuring the wind for primarily weather forecast purposes. These data have sometimes the right properties for wind assessment studies and sometimes not, that is why it is of importance to be critical to these types of data. Simulated data may also have different purposes than describing the wind speed at hub height.

2.3.1 Measurement mast on sites

The most reliable source of data to calculate the expected wind potential is an onsite wind measuring mast or similar worthy measurement equipment that has been measuring the wind speed, wind direction and temperature for at least one year. This is normally not the case, and data from a short wind measurement period has to be used and transformed via models that can estimate the long-term wind environment. The measurement equipment also has to be free from disturbance and long-term failure to be usable for wind estimation. To study the statistical parameters of wind estimation, the measuring series have to be long enough to cover the yearly variations, that is, all the seasons on a site.

2.3.2 Normal year correction

There are several mathematical models that are capable of doing an estimation of the wind speed at sites with the use of short-term measurement series on site correlated with long-term measurement series from a position near the site. The most common ones are the Measure-Correlate-Predict method and the Wind Index method. These two methods differ a great deal from each other, due to the different sources of data used. The Measure-Correlate-Predict method uses actual wind speed measurements and tries to correlate the measured data with a reference series from a nearby meteorological station or another source of long-term data. Via the relationship between these two series it is possible to get a long-term corrected wind speed series calculated for the site (Burton et al., 2001). The Wind Index method, on the other hand, tries to estimate the wind speed from the wind turbine effect and its energy production. To be able to use this method well, a normal year correction has to be performed on the energy production data so that the produced energy can be compared over time. This method does not consider the wind direction, which can be useful in e.g. complex terrains.

2.4 EARLIER DISCOVERIES IN LONG-TERM CORRECTION SURVEYS

Since the need for long-term corrected data is of great interest for the wind power industry, a lot of time, research and analysis have been put into determining the accuracy of methods covering this area. On the current wind assessment market there are a few alternatives to choose from regarding computer programs describing and calculating wind energy content, e.g. WindPRO (EMD International, 2009), openWind (AWS Truewind, 2008), Windfarm (WindFarm , 2009) and Windfarmer (GH WindFarmer, 2009). All of these programs have modules for doing MCP and energy content calculations, and as the names reveal, they are also able to do wind farm calculations and optimizations. They come from four different companies with solid

background in the wind power business: EMD, AWS Truewind, ReSoft and Garrad Hassan. In this thesis calculations will be computed with WindPRO due to its availability at Vattenfall Wind Power.

There is a discussion that fundamentally splits the views of modeling wind energy content in two parts. The question is how the accuracy and usage of global models as a reference series in long-term correction of wind data can be considered good. There is one study by AWS Truewind "The use of reanalysis data for climate adjustments" (AWS Truewind, 2006), which points out that Reanalysis data are unpredictable when used in MCP methods. In this article the writer uses the term Reanalysis data for data reanalyzed from the period between the years 1948 till 2006. It is shown that in a comparison with a rawinsonde (radio wind sonding) at a pressure level of 700 hekto Pascal (hPa) in Denver, Colorado, the Reanalysis data have a downward trend over a 30-year period. The Reanalysis data had a mean wind speed 23 % higher in the beginning of the series and only 9 % higher in the end of the series. This is considered strange by the author, since the rawinsonde in question is used as an input parameter in the Reanalysis data. The most likely reason for this inconsistency is that so many other data sources are added that the model becomes imprecise. In Hemsby, Great Britain, another rawinsonde was compared with Reanalysis data at a pressure height of 850 hPa. This gave, in contrast to the first comparison, not a trend but a year-to-year inconsistency. Here there are no major mountains or obstacles present, as in the first study, but the two series follow each other between the years 1973 and 1977. After these years, in 1979, other sources than rawinsondes were added as input sources in the Reanalysis data series, e.g. airplane and satellite data. The final conclusion of the report is that the Reanalysis data cannot be preferred over real rawinsonde data. It is also pointed out that extending Reanalysis data further back than 10 years can result in fluctuations and an increased risk for errors (AWS Truewind, 2006).

In the report "Från mätt vind till vindklimat" by Erik Nilsson and Hans Bergström (Nilsson & Bergström, 2009) written in the Elforsk program, a concrete comparison between geostrophic wind and Reanalysis-2 data was evaluated. The purpose of the report was to establish if the geostrophic wind at a pressure height of 850 hPa can be used as a long-term reference series in normal year correction. One reason that supports this theory is that the geostrophic wind is what drives the actual wind at e.g. hub height. If it is possible to estimate the wind resources at a site through the geostrophic wind and pressure measurements, it would reduce the complexity in performing adequate wind measurements. It is likely that the air pressure measurements are less inhomogeneous than surface wind speed measurements. Three different methods are compared, two linear and one of a higher order. The results are compared against not doing any normal year correction.

The concluding part of the report states that measurement series less than a year long, have an increased risk for errors. When using one year of measurement data from the measuring mast Näsudden, at 75 meters' height, there is a 5 % risk of incorrectly estimating the wind speed with 0.42 m/s. If another year of measurements is added, the 5 % risk of incorrectly estimating the wind speed decreases to 0.32 m/s. In the case

where no method is applied to correct the measurement series, the 5 % risk of incorrectly estimating the wind speed is 1.0 meters per second (Nilsson & Bergström, 2009).

One question raised by Vattenfall Wind Power was if there were any significant variations in the resulting deviation when performing a wind speed measurement in different seasons. If there seems to be a deviation, how does it deviate? Does it over or underestimate the wind speed during different seasons? It is already known that the wind varies during the year with higher wind speeds in winter periods and lower wind speeds in summer periods. To conclude, if the seasonal dependence of the measurement continues over time, the MCP methods provide useful information.

The discussions are not just about the data assimilation methods, they are also about how to, and with which method, the data should be treated. Available in the statistical community is a massive storehouse of statistical methods for doing different resample and regression methods on wind data. Although there is some general agreement in the discussion, there is still some debate in how to determine which method is of most interest for different sites.

3 THEORY

3.1 WIND ENERGY CONTENT

It has always been of interest for wind power engineers to be able to evaluate meteorological parameters such as wind speed, geostrophic wind and wind climate on site to be able to give as correct an assessment of the wind potential, and thereby the power production, as possible.

In wind energy estimation the mean wind speed at hub height is of special interest. Hub height is defined as the height above ground of the turbine. Shown in Equation 1, there is a relationship between the wind speed, v m/s, the air density, ρ kg/m^3, the area swiped by the rotor, A m^2 and the kinetic wind energy, $P_{kinetic}$ W/m^2 (Wizelius, 2007).

$$P_{kinetic} = \frac{1}{2}\rho A v^3 \quad \left[\frac{\mathrm{W}}{\mathrm{m^2}}\right] \tag{1}$$

According to Equation 2 there is also a limit to how much of the potential wind energy a wind turbine is able to use. Betz's Law gives the relationship between kinetic wind energy and theoretical maximum energy that a wind turbine can obtain from the wind. Wind power turbines can only assimilate about 59 % of the actual kinetic wind energy (Wizelius, 2007).

$$P_{actual} = \frac{16}{27} P_{kinetic} \quad \left[\frac{\mathrm{W}}{\mathrm{m^2}}\right] \tag{2}$$

When the wind speed and the wind distribution are known, it is possible to calculate the expected production from a specific wind turbine. This is possible to do with a power

output curve which is specific for each type of wind turbine and states the power output from the turbine at different wind speeds. A small error in the wind assessment process can result in significant deviations in the expected wind energy production.

3.2 WIND DISTRIBUTION

The Weibull distribution is closely related to the exponential distribution. This distribution has two main parameters in terms of wind speed assessment, the shape factor, c, and the scale factor, k (Johnson, 2005). In the wind speed distribution there is a higher frequency of lower wind speeds than the mean wind speed. This results in a curve with a peak on the lower end of the distribution and a tail on the higher end. From the peak the shape factor can be determined. It is the shape factor that determines how stable the mean wind speed is at a site. If the shape factor is equal to 1, the distribution is called the exponential distribution; if equal to 2, the Rayleigh distribution, and if equal to or greater than 3, it converges towards the Gaussian distribution. Normally a value of the shape factor around 2 is assumed for wind distribution, i.e. following the Rayleigh distribution (Patel, 2006).

From the probability density function for a Weibull distribution, Equation 3, the mean value and the variance can be calculated.

$$f(x) = \begin{cases} kcx^{c-1}e^{-kx^c} & \text{for } x > 0, k > 0, c > 0 \\ 0 & \text{elsewhere} \end{cases} \quad (3)$$

The mean value, μ, can be calculated from the integral shown in Equation 4.

$$\mu = \int_0^\infty x\, kcx^{c-1}e^{-kx^c}dx \quad (4)$$

The variable transformation $u = kx^c$ transforms the integral in Equation 4 to an expression with the shape parameter and the scale parameter in combination with the gamma function, Γ, as seen in Equation 5.

$$\mu = k^{-\frac{1}{c}}\int_0^\infty u^{-\frac{1}{c}}e^{-u}\,du = k^{-\frac{1}{c}}\Gamma\left(1+\frac{1}{c}\right) \quad (5)$$

The variance of a Weibull distribution, σ_W^2, can be expressed using the same parameters and combining them with the gamma function, as seen in Equation 6.

$$\sigma_W^2 = k^{-\frac{1}{c}}\left\{\Gamma\left(1+\frac{1}{c}\right) - \left[\Gamma\left(1+\frac{1}{c}\right)\right]^2\right\} \quad (6)$$

How the mean wind speed deviates when the Weibull shape and scale parameters change is described in Figure 1Figures 1 and 2. These graphs were generated using the MATLAB Weibull random distribution function to generate series with different properties regarding scale and shape parameters. The skewness, describing the shift of the curve, and the kurtosis, describing the peak of the curve, are also stated for each graph.

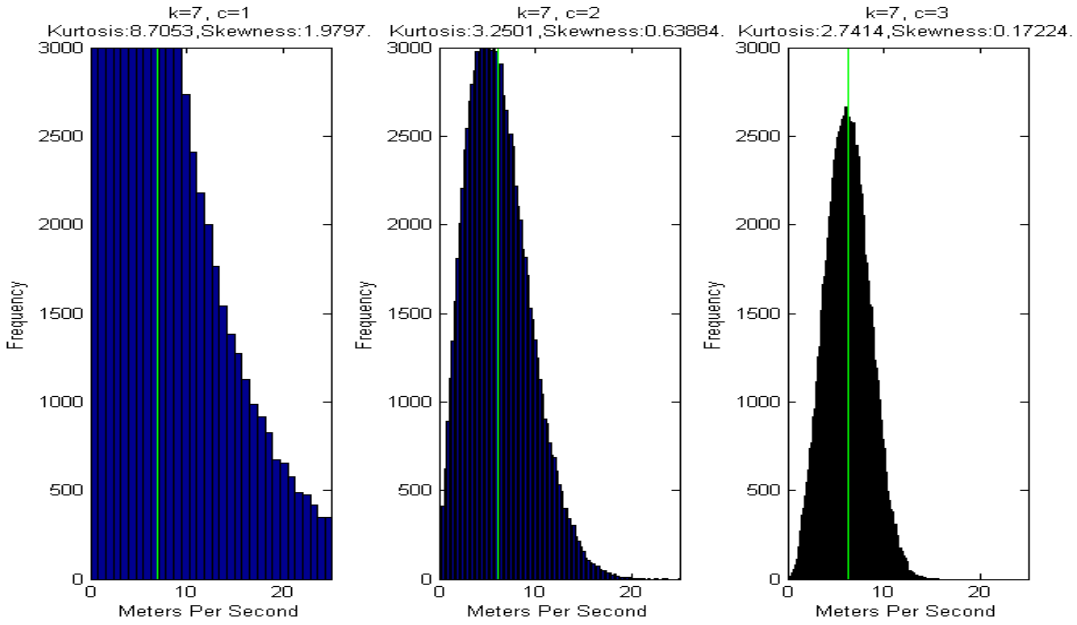

Figure 1. Weibull distribution plots with changing shape factor (c) and constant scale factor (k).

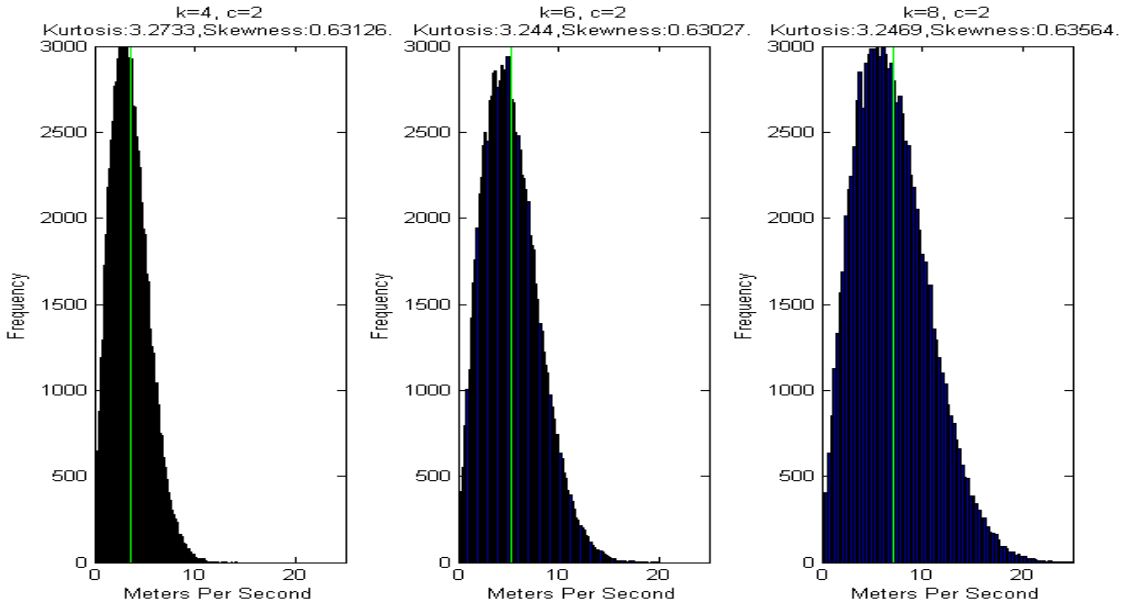

Figure 2. Weibull distribution plots with changing scale factor (k) and constant shape factor (c).

3.3 GLOBAL MODELS

To be able to use the Measure-Correlate-Predict method there has to be one measurement series and one reference series. As described above there are data series that are measured with real measuring masts covering extensive areas and there are

computer models that simulate data series over the whole world in virtual grid points. There is a quite hands-on approach using the measured data series from real masts. With the series simulated by computers, there are different approaches. In this study there are two kinds of data series, the NCAR/NCEP Reanalysis-2 data series and the Other Meteorological Data (Ackerman & Knox, 2003).

3.3.1 Estimation limitations

In many wind calculation programs data are only used from some of the levels of the Reanalysis-2, e.g. wind speed at a simulated height of ten meters above ground level or the wind speed at one of the pressure heights or a sigma level. The difference between sigma levels and pressure levels is that the sigma levels correspond to an actual height above ground. Hybrid coordinates that are defined by height above ground make this possible. The advantage with using a sigma level instead of a pressure level is that a pressure level shifts with its height above ground, and for a low pressure level near the surface pressure, it can sometimes be situated under the ground level. The sigma level is always at the same height above ground and defined as the pressure at the sigma level divided with the surface pressure. This results in a value between one and zero, with zero being in the top of the atmosphere and one at the surface level. This can be a reason for precaution when comparing the measuring masts with Reanalysis-2 data. The measuring mast and the Renanalysis-2 do not have the same heights. The fluctuations of the pressure levels decrease with height, so at the pressure level 850 hekto Pascal (hPa) the effect of the topography is reduced. This is also true for the sigma levels.

Using data with high resolution in the nodes, but with a large grid, will result in big gaps of knowledge in the meteorological environment in the nodes. The Reanalysis-2 data do not have the ability to correct for small changes in the wind because there is the grid size of 2.5° latitude and 2.5° longitude spread over the entire earth's surface. To be able to judge critically the uncertainty of the wind resources and determine if a site is suitable for wind power production, there has to be a high correlation between the measuring mast and the Reanalysis-2 data (Kistler et al., 1999).

3.4 THE MEASURE-CORRELATE-PREDICT METHOD

When an on-site measuring mast or a nearby normal-year-corrected data series is available, the Measure-Correlate-Predict (MCP) method can be used. There are different types of MCP methods that have different statistical focus and thereby can be used in different situations. In long-term correction of wind data, methods like Regression, Weibull scale, Matrix or Wind index MCP can be used (Thøgersen et al., Undated). In this thesis only MCP regression methods will be studied.

In addition, which parameters can quantify and validate methods used in long-term correction will be studied.

3.4.1 Measure-Correlate-Predict estimation procedure

All regression methods are based on the same theory. Regression methods assume that wind speeds from different sources have a linear or curved relationship to each other. It is possible to relate a reference series against a measured series and make a linear or

curved estimation line run through the data points. The resulting equation describes how the reference series vary in scale and offset to the measured series. This procedure makes it possible to estimate the wind speed at a site with only a short measuring period if a long reference series is available. This is a widely used method in wind assessment estimation today. MCP methods have been proven to work well at many sites with uncomplex terrains, but when applied in highly forested areas or complex terrain there are obstacles that have to be considered and investigated in order to use the method correctly.

3.5 STATISTICAL METHODS FOR VALIDATING WIND ESTIMATIONS

To be able to minimize the uncertainty of a wind potential study, there is a storehouse of statistical tools that can be of interest to get a better estimation of the data. When quantifying the uncertainty of the wind potential at a site, the prediction of the wind velocity is crucial. In order to succeed with a long-term estimation of wind speed there has to be a solid reference series.

3.5.1 Data validation

Measurements by a measurement mast are not always technically valid and cannot be used without careful validation. If errors occur in the data set, there will be continuous bias in the treatment of the data. First of all there is always the possibility graphically to validate the data series used, normally by plotting the data against time. For example, passages with zeros will reveal themselves. If there is a bias, e.g. a change in the measuring equipment resulting in a different mean wind speed, this will also be noticed.

Cook's distance is used to detect and quantify outliers in a data set. An outlier is defined as a data point which is unrepresentative for the data set as a whole. When calculating the Cook's distance, the residuals and the unusualness, also called leverage, of the data points are taken into account. The residuals are the difference between the measured value and the model value. Data points with high residual values or great unusualness, or both, are given a high Cook's distance. If the Cook's distance is greater than one for a data point, it is assumed to be in need of investigation (Heiberger & Holland, 2004). The calculation of the Cook's distance for the i-th data point is shown in Equation 7, where e_i^2 is the squared residual for the i-th data point, p is the number of unknown parameters, h_i is the leverage for the i-th data point and E_{MS} is the mean square error (The MathWorks, 2005a).

$$D_i = \frac{e_i^2}{p\, E_{MS}} \left(\frac{h_i}{(1-h_i)^2} \right) \qquad (7)$$

3.5.2 Descriptive statistics

Parameters that are used to get an initial overview of the data series are the mean, the median, the variance, the standard deviation, the skewness and the kurtosis and the

correlation coefficient between series. A brief summary of the theory behind these parameters is given below.

The mean is the sum of all the values,X_i, in the data series divided with the total number of data points of the series, N (The MathWorks, 2005b). It is called the arithmetic mean.

$$\bar{X}=\frac{1}{N}\sum_{i=1}^{N} X_i \quad (8)$$

The variance is a measurement of how much the data series disperse (The MathWorks, 2005c). The unit for the variance is always a squared factor of the measured value.

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2 \quad (9)$$

The standard deviation, with the same unit as the measured data series, describes the dispersion from the mean as well but in actual units (The MathWorks, 2005d).

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2} \quad (10)$$

The covariance is calculated because it is needed when calculating the correlation coefficient (The MathWorks, 2005e).

$$Cov(X,Y) = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y}) \quad (11)$$

The correlation coefficient, $R(i,j)$, describes the correlation between the data series, where values between minus one, a perfect negative correlation, and one, a perfect positive correlation, are possible (The MathWorks, 2005f).

$$R(X,Y) = \frac{\sum_{i=1}^{N}(X_i-\bar{X})(Y_i-\bar{Y})}{(N-1)\sigma_x\sigma_y} \quad (12)$$

The behavior of the frequency distribution of a data series can be studied by calculating the skewness. It is used for describing the asymmetry of a distribution. If the value of the skewness is negative the distribution is skewed to the left, and if the value of the skewness is positive the distribution is skewed to the right. A symmetric distribution has a skewness value of zero (The MathWorks, 2005g).

$$Skewness = \sum_{i=1}^{N} \frac{(X_i - \bar{X})^3}{\sigma^3} \quad (13)$$

The kurtosis value can be used to estimate how the distribution varies in regards to the normal distribution. If the kurtosis value is greater than three the distribution is more peaked than the normal distribution, and if less than three the distribution is flatter than the normal distribution (The MathWorks, 2005h).

$$Kurtosis = \sum_{i=1}^{N} \frac{(X_i - \bar{X})^4}{\sigma^4} \quad (14)$$

3.5.3 Linear and curve estimations

One of the oldest and most common ways of describing the difference of time series is through the Ordinary-Least-Square regression method (OLS); see Equation 15. Due to its easily understood theory, it has become an estimation used in various situations and it is widely accepted. Even though the OLS regression method is sensitive to outliers the use of this method is widespread. Investigating Equation 15 it is noticeable that the OLS regression method should be used with caution due to the squaring of the difference between the estimated and the true value. In cases where the data are uniformly centered on a clear linear line, the OLS regression method is the obvious choice. But when outliers are suspected to be part of the data, there are alternative estimations that are less affected by outliers than the OLS regression method (Good & Hardin, 2006). Y_i represents the dependent data set, a where the regression line intercepts the y-axis, b the slope of the regression line and X_i the measured data set.

$$\sum_{i=1}^{N} (Y_i - a - bX_i)^2 \quad (15)$$

Instead of using a method that squares differences between values and the regression line, a method that takes the absolute value of the differences between the regression line and the values may be implemented for testing; see Equation 16. The resulting regression curve is then less sensitive to outliers, as shown in Figure 3. Instead of minimizing the sum of the linear equation it tries to minimize the sum of the absolute deviation. It results in a more stable regression method, which is less sensitive to outliers. This method is called Least-Absolute-Deviation regression (LAD) (Good & Hardin, 2006).

$$\sum_{i=1}^{N} |Y_i - a - bX_i| \quad (16)$$

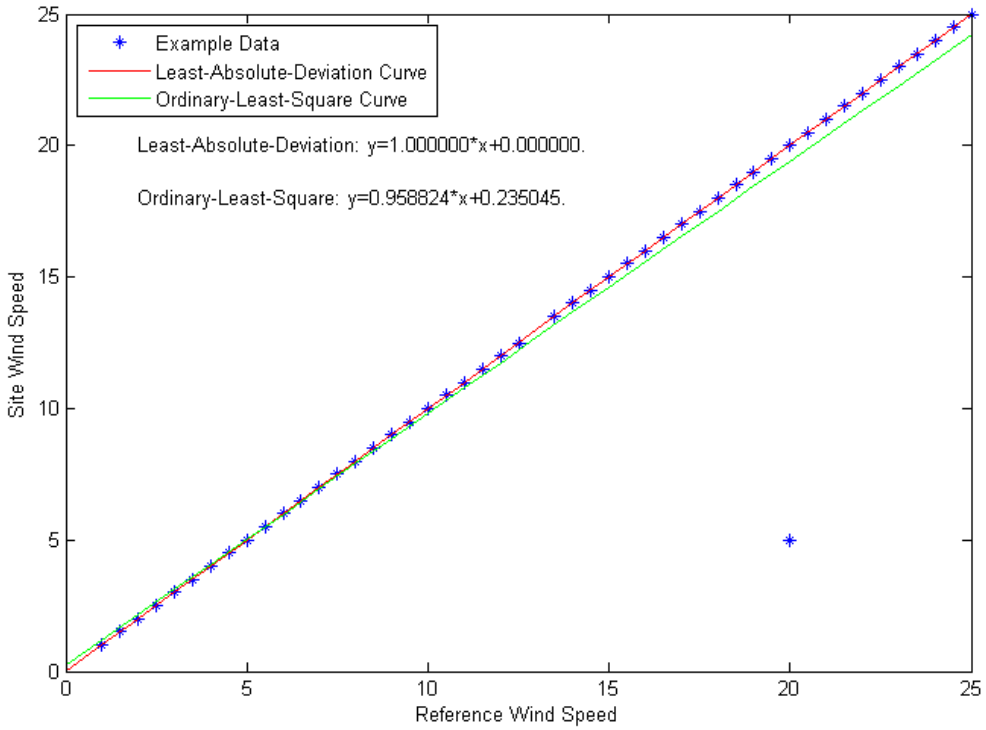

Figure 3. Ordinary-Least-Square regression curve versus Least-Absolute-Deviation regression curve with one outlier.

One method proven to be useful in validating the accuracy of a linear regression fit in terms of mean and standard deviation of a data series is the cross validation method. It parts the *n* data points into *n-1* pieces of the same size, using 1 piece for validation. The model is run *n-1* times, each time with a part of the data blocked out, e.g. on the first run only the first data set is blocked out, on the second run only the second data set is blocked out and the first data set from the first run is reinstated. This results in *n-1* linear regression lines, each describing the data set without a part of the data available. Through this method it can easily be detected if some parts of the data are abnormal. It is also possible to run the model for all the data points, creating *n-1* linear regression curves, which gives a good graphical overview of which points contribute especially to the variations (Trauth, 2006).

To investigate if the data have a correlation between some higher polynomial regressions, e.g. Curve Linear regression, Equation 17 can be used. In cases where the data are described with a linear regression curve, the coefficient in front of the higher order polynomials normally is much smaller than the coefficient in front of the first order polynomial. This means that the polynomial regression curve has a small second order influence and it should not be used to describe the data (Trauth, 2006).

$$Y_i = \mathrm{b}_0 + \mathrm{b}_1 X_i + \mathrm{b}_2 {X_i}^2 \qquad (17)$$

When it is of interest to consider the errors in both data series, the Reduced-Major-Axis (RMA) regression method, see Equations 18 and 19, can be considered (Sinclair &

Blackwell, 2002). The RMA regression method minimizes the area between the data points and the regression line, by using the difference between the estimated and the measured values of the data series. This is a complex optimization, which has proven to be describable via the standard deviation and the mean value of the data series, thereby reducing the complexity significantly (Sinclair & Blackwell, 2002). The slope, b_1, can be calculated from Equation 18, and the y-intercept, b_0, from Equation 19 (Trauth, 2006). Then b_0 and b_1 are used to describe a linear regression curve of the same form as OLS and LAD.

$$\mathrm{b}_1 = \frac{\sigma_y}{\sigma_x} \tag{18}$$

$$\mathrm{b}_0 = \bar{Y} - \mathrm{b}_1\bar{X} \tag{19}$$

3.5.4 Resample estimations

To validate the equation given by a regression model, resampling methods are of interest. When a resampling method is applied to a data series, the data are resampled in some way, thereby describing the variability and the precision of the selected variable, e.g. slope, the y-intercept, mean value or the standard deviation.

Resampling methods were discovered long before there were computers with today's capacity. This is why the resampling methods have become more interesting in areas where the predicted variable is assumed to have a bias and some variance. It is nowadays possible to do enormous resampling estimations in a short time, which gives a result that fairly easily can be used as a validation tool for statistical parameters.

There are two resampling estimations that have spread to broad usage in different scientific areas that will be explained and used here. These are the Jackknife and the Bootstrap estimations. The Jackknife estimation was introduced in 1949 by Quenouille, and the Bootstrap estimation was introduced in 1979 by Efron (Shao & Tu, 1995).

The Jackknife estimation requires a certain smoothness level of the used data series. If this is not the case, an adjustment to the Jackknife estimation needs to be done. This adjustment basically divides the data series into subsamples. This is much like the cross-validation method. If the smoothness of the data series is low, the subsamples have to be bigger to get a valid estimation. The adjusted Jackknife method is called the Deleted-d Jackknife estimation (Shao & Tu, 1995).

The other method, Bootstrap estimation, often gives a result near the result of the Jackknife estimation, but the Bootstrap uses a much more complex resampling method. Bootstrap estimators are a mixture of two types of resampling methods, partly a substitution method and partly a numerical approximation method. The numerical method, which is based on the Monte Carlo approximation, is more relevant in practical problems. Due to its complexity there are many developments of the Bootstrap estimator, which has resulted in a variety of expansions (Shao & Tu, 1995).

There are many differences between the Jackknife method and the Bootstrap method. When the distribution of the data series is heavily tailed, the Bootstrap may not be a good choice due to inconsistency in the resampling populations. The Jackknife on the other hand is capable of giving consistent results even when the distribution is tailed, but there has to be a certain smoothness level of the data set. The Jackknife estimation is a more robust method for making assumptions of a statistic than the Bootstrap estimation. It is also less complex, which makes it possible to easily extend the Jackknife into more difficult estimations, e.g. multivariate cases. The Bootstrap estimator is normally not recommended if only simple estimators are needed (Shao & Tu, 1995).

These two methods describe the same thing but use different calculation procedures. This is why it is interesting to see the difference in appearance of these two resampling methods. Also, they are relatively easily applied in comparison with the Monte Carlo estimation, which is another well-known experimental method for estimating probability distributions for a chosen function (Gentle, 2009).

If a linear regression has been applied, it is useful to know what correlation the data points have. If so, the correlation coefficient shown in Equation 12, R or also noted *r*, should be examined. The correlation coefficient is a measurement of how good the relationship is between two data series (Urdan, 2005). This parameter can also be resampled and described by both the Jackknife and the Bootstrap estimation method.

3.5.5 Empirical and spectral analysis estimations

When determining the goodness-of-fit of the regression models for two data series, not only theoretical procedures are available. It is also possible to make empirical assumptions and estimations from the results of the regression model. Basically, an empirical estimation tries to find a relationship in measurement data that does not have to be explained, or sometimes cannot be explained by theory. One empirical method often used, is to take the cumulative value of the statistic that is of interest, and plot it against time. It can be established whether the statistic has a tendency of stabilizing for some value.

Parameters of interest for an empirical estimation can be, as mentioned in the presentation of the resampling estimation above, slope, y-intercept, mean, median, standard deviation and kurtosis. This might give some understanding of how long and in what resolution of time, a measurement of the parameter should continue. If the data series available are good enough, the parameter can be impaled with a chosen standard deviation and thereby quantify how good the measurement is at that point. Knowing when a measurement series can be classified as long enough in terms of usage for a specific purpose is important in both a statistical and an economic sense.

4 METHODS FOR WIND DATA ESTIMATION

There are two types of data available in this study, measured data from measuring masts and virtually simulated data.

4.1 AVAILABLE DATA

In this study the available data are from measuring masts at the nuclear power plants Ringhals and Oskarshamn and from a measuring mast located north of Uppsala, Marsta; see Figure 4. As the long-term reference measurement series there are wind speed data from the NCEP/DOE AMIP-II Reanalysis project and the Other Meteorological Data available.

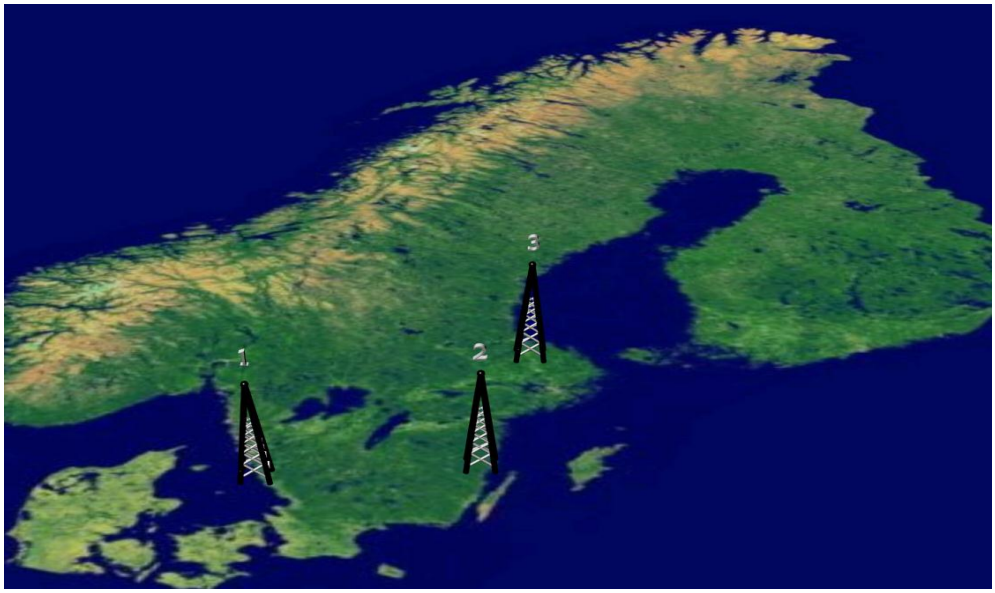

Figure 4. Map over Sweden showing the three measuring masts, 1 Ringhals, 2 Oskarshamn and 3 Marsta (Map from Hitta.se, 2009).

4.1.1 Measuring masts

The measuring mast situated a little south of the nuclear power plant Ringhals has three measuring heights, 24, 48 and 96 meters, where wind speed and directional data are collected and averaged every five minutes. The mast is situated on the west coast of Sweden. Usable data from nine years in between 1996 and 2007 have been retrieved from this measuring mast. The position of the mast is expedient for studying purposes because of the close distance to a NCAR/NCEP Reanalysis-2 node. One Reanalysis-2 node is situated right on the measuring mast.

The measuring mast outside of Oskarshamn is also a near-coast mast on the east coast, measuring wind speed and directional data at the heights 25, 75 and 100 meters. Data from eight years of consistent measurement were available from this measuring mast, between 2001 and 2008. The mast is located between two Reanalysis-2 nodes, which make it difficult to use more than those due to the long distance to the next nodes. Between the onshore node and the offshore node there is a coastline and two islands. This is a very complex situation for the Reanalysis-2 model to give accurate values because the model grid has higher resolution than the two islands together. The node

situated west of the measuring mast has been chosen due to a better correlation coefficient than the measuring mast.

In Marsta, situated 8 km north of Uppsala, data from four years of measurement have been retrieved from a 29-meter high measuring mast. Measurements were performed between the years 1994 and 1998. This is a short measuring mast but it can reveal interesting facts about regression methods as well as the reference series. There is a Reanalysis-2 node situated almost on this position, which makes it possible to use data from the 850 hPa level and the 42-meter sigma level.

4.1.2 NCAR/NCEP Reanalysis-2

The data series that have been extracted from the Reanalysis-2 project are on the pressure level 850 mb. From WindPRO a data set from the sigma level 0.995 has been extracted, which corresponds to a constant height of 42 meters above ground. These two data series are available for all Reanalysis-2 nodes used here and extend from 1979 to 2008.

4.1.3 Other Meteorological Data

The content is excluded due to proprietary material. This content in the official version is called Other Data or Other Meteorological Data.

4.2 DATA VALIDATION

A first visual validation of the mast data was easily made by simply plotting the data against time. It gave a relatively clear picture of availability of the measurement series. Also gaps and places where the measurement equipment generates a constant to signal an error can be seen. Examples of this type of behavior are values when the anemometers had frozen and given the value zero, or when gaps occurred in the date column due to electrical failure. These values were eliminated in the measurement series, as were the corresponding values in the reference series. When the first validation was satisfactory, a second validation began. The second validation was applied to correct the data series in time, i.e. to shift the series towards correct time stamps. This problem occurs when there has been a gap in the time stamp for some reason. The shift in the indexing of the measurement series regarding the reference series is corrected by comparing every time stamp from the measurement series with the reference series. If there is a gap in the measurement series, the corresponding value in the reference series was removed.

One further complication regarding the time stamps in the Reanalysis-2 data was discovered during this procedure. The format of the time stamp in which NCEP/NCAR stores the Reanalysis-2 data is by hours since midnight of January 1, 1800. It might seem strange to use this kind of time stamp, but with this method there is no correction needed for leap-years. A script was built to transform the seven-digit hour value into year, month and day so that the Reanalysis-2 data were more easily compared with other data sources. During the time shift procedure a graphical control could be made. A plot was created showing the day stamps of the two time series and how they successively overlapped each other.

4.3 CALCULATION PROCEDURES

The initial step in the calculation procedure was to convert all data into the same structure. The important parameters are the wind speed, the wind direction and the time stamp. All the Reanalysis-2 data are stored in a format called NetCDF, .nc-files, which are four-dimensional storage files with longitudinal, latitudinal and vertical coordinates corresponding to time. In MATLAB the relevant data have been extracted through a toolbox, named MexNC-toolbox (Evans, 2009). This toolbox consists of a .m-file package that transforms .nc-files into MATLAB standard matrixes. It is possible to manipulate the data and scale the grid down to a more relevant size, e.g. specific nodes in the Reanalysis-2 data series. The Reanalysis-2 data are given in u and v-components, i.e. one component containing the east-west direction and one the north-south direction. The directional data can be achieved by taking the value of the tangents for all the values in the north-south component divided with the values in the east-west. When this was done daily, means could be calculated and stored into standard structure matrices for the Reanalysis-2 data and all of the measurement masts.

4.4 IMPLEMENTING STATISTICAL MODELS ON WIND DATA

When the data series had the same structure and the time stamps were corrected it was possible to start testing the wind data series. Numerous methods and statistical tests were implemented via functions created in MATLAB. As described in section 3.5 these methods can be divided into different groups, regression methods, resample methods, empirical methods and also methods showing the residuals' distribution.

The linear regression method OLS, which is one of the most commonly used methods for correlating two data sets, is easily implemented in MATLAB. There is a backslash operator that performs a least square solution between two matrices. The only modification needed is adding a column of ones before the measuring mast matrix and the backslash operator will return the coefficients for the least-mean-square curve. To estimate the LAD regression line, on the other hand, is a much more complicated procedure in MATLAB due to the feedback calculations needed. The residuals from the last iteration are used to calculate the new weight function, which in turn is used to calculate the new coefficients. When the difference between the old and the new coefficients is smaller than a chosen value, the calculations end. This value was here chosen to be 10^{-6} and the iterative value was chosen to be 10^{-5}. The choice of these two variables was based on the assumption that they were small enough; if proven wrong they would have been easily corrected. As seen in section 3.5.3, the RMA regression method and the Curve Linear regression method are both relatively easy and uncomplicated to implement in a calculation program like MATLAB because of the preexisting commands for this type of calculations. Calculating the cross validation method demands a bit more understanding of the statistical procedure. The difference between the cross validation method and the Jackknife estimation is not that great. In the cross validation method there is one section of data removed for every calculation; in the Jackknife estimation there is only one value removed at a time.

The Jackknife, mentioned briefly above, and the Bootstrap method are resampling estimations. In MATLAB there is already a Bootstrap function where the number of data samples and the function used to do the computations can be chosen. A greater number of resampling gives a clearer picture of the result but demands more computer power. Here the number of data samples has been set to 1000, which gives a good accuracy and an affordable simulation time. The Bootstrap estimation, as the Jackknife estimation, can be used to describe different kinds of statistics via different choices of functions controlling the method. Functions were chosen to describe the correlation between the measured data series and the reference data series. It was also used to estimate the mean and standard deviation of the y-intercept and the slope of the regression curve. There is a Jackknife estimator prebuilt into MATLAB, but since this method is simple to compute and thereby control, the study's own jackknife function was created. A loop is stated to run through all of the samples and estimate a regression line of the first order. The coefficients are stored and analyzed when the loop stops.

To be able to validate the methods, residuals were plotted. Residuals are defined as the model value subtracted from the measured value. The reference series subtracted from the measured values results in the residuals. Residual plots can show how big the variations are between the series and also if there are any trends. If trends occur in the residuals it means that the model used to describe the data suffers from a bias.

All the above-mentioned and explained methods and estimations are useful when the true values are hidden in a data series. Luckily in this case the measurement series are long enough to do adequate cumulative empirical estimations of almost all of the parameters in the regression and distribution methods. This means that the value for a statistic can be calculated cumulatively throughout the process, i.e. first just the first two points in the data series are used as input data, after that the first three points and so on. A cumulative empirical calculation of this kind results in, if the data series are long enough, a curve that converges. So in the case with the OLS regression, the y-intercept and the slope are calculated one time less than the length of the data series due to at least two values being necessary to do a linear regression. In a plot of these two parameters against time it is clear that they converge to the same value as the calculation of the regression line. What also is shown is the time it takes for the parameters to stabilize. The standard deviation was also added in the graph to show the improvements for each time step. These graphs are not included in the report because they were a test during the calculations for immediate checking of the behavior of the parameters.

A time series of wind speed is as mentioned in section 2.3 assumed to follow the Weibull distribution. To compare if the reference series and the measurement series have the same distribution, the Weibull distribution function in MATLAB is used. This function will return the shape and the scale factors for the data set inserted. These parameters can be compared for two data sets and should be close to each other. Another method to see how the distributions behave is to do a kurtosis calculation of the data sets. This is a method that uses the mean and the standard deviation of the distribution and calculates a value that estimates how great the peak of the data set is. If

the kurtosis is equal to three, the distribution follows the normal distribution; if the value is less than three, it means that the data set has a smaller peak than the normal distribution and if greater than three, a bigger peak.

Calculating the deviation from measured mean wind speed is here stated by dividing the model mean wind speed with the measured mean wind speed and multiplied with one hundred. The deviation is stated in relationship to 100.00 % which makes it easier to compare different regression methods and reference series with each other, as can be seen in Equation 20.

$$Deviation, \text{comparable} = \frac{\hat{\bar{X}}}{\bar{X}} 100 \quad (20)$$

In the part where the length of the measurement data and the seasonality are investigated, the percentage value has been subtracted with 100.00 which makes it possible to see deviations in a smaller scale. In the case where the measurement length is considered the absolute value of the deviations has been used; see Equation 21. This is because the focus on deviation is greater than if an over or underestimation occurs. In the seasonal case the absolute value has not been used because here the focus has been to see if over or underestimations have occurred; see Equation 22.

$$Deviation\ in\ percentage = \left| 100 - \frac{\hat{\bar{X}}}{\bar{X}} 100 \right| \quad (21)$$

$$Deviation\ in\ percentage = 100 - \frac{\hat{\bar{X}}}{\bar{X}} 100 \quad (22)$$

4.5 ERROR ESTIMATION OF THE MODEL

As in all models it is crucial to define the errors and flaws the models have. Before adding the measured wind data into the model, all data points with wind speeds equal to zero were removed. The reason for this is that when anemometers freeze they cannot move, hence not measure the wind speed, and consequently return the value zero. Assumptions have been made that the few five or ten minute mean values that actually have a value of zero meters per second are extremely few compared to those hours and days the anemometers freeze. It can therefore be assumed that when a five or ten minute mean value of the wind speed is zero, there is something wrong with the measurement. When the daily mean values were calculated, there were days with incomplete date notation due to electrical failure and due to the correction of values equal to zero. When this occurred the mean value was calculated with the remaining values. This is a risky maneuver because of the differences in time of the day. The wind speed differs a great deal in coastal areas, where some of these measurement masts are situated, due to the sea breezes. For example, if one day only has the first six hours available, the mean

value of the wind speed and the wind direction will be based on unrepresentative data. This will give a bad correlation with a global model because the global model will have a full daily cycle with which to calculate its mean wind speed and wind direction. The occurrence of this type of behavior in combination with missing data points is considered below 2 % for all the measured data used.

5 RESULTS

To minimize confusion in how much of the reference series is used, the amount is stated inside brackets, e.g. (9 y) means a nine-year data series has been used.

5.1 DESCRIPTIVE RESULTS

For the explanation behind the parameters in Table 1, Table 2 and 3 see section 3.5.2 Descriptive Statistics.

Table 1. Wind data from Ringhals measuring mast at height 96 meters above ground for nine of the years between 1996 and 2007 and corresponding years from the reference series.

						Cross validation	
Wind data series	$\bar{X}$ [m/s]	σ [m/s]	Kurtosis [-]	Skewness [-]	r [-]	Error of $\bar{X}$ [m/s]	Error of σ [m/s]
Ringhals 96 m (9 y)							
850 hPa (9 y)	8.31	4.29	3.05	0.58	0.82	-0.24	2.90
42 m Sigma (9 y)	6.22	2.84	3.71	0.74	0.80	1.70	3.23
Other Data (9 y)	7.94	3.53	3.08	0.57	0.93	0.005	3.15

Table 2. Wind data from Oskarshamn measuring mast at height 100 meters above ground from the years 2001-2008 and corresponding years from the reference series.

						Cross validation	
Wind data series	$\bar{X}$ [m/s]	σ [m/s]	Kurtosis [-]	Skewness [-]	r [-]	Error of $\bar{X}$ [m/s]	Error of σ [m/s]
Oskarshamn 100 m (8 y)							
850 hPa (8 y)	8.40	4.38	3.12	0.62	0.73	-0.62	2.10
42 m Sigma (8 y)	6.89	3.10	3.28	0.62	0.79	-0.01	2.09
Other Data (8 y)	7.92	3.37	2.98	0.52	0.87	-0.66	2.09

Table 3. Wind data from Marsta measuring mast at height 29 meters above ground from the years 1995-1998 and corresponding years from the reference series.

Wind data series	$\bar{X}$ [m/s]	σ [m/s]	Kurtosis [-]	Skewness [-]	r [-]	Cross validation Error of $\bar{X}$ [m/s]	Cross validation Error of σ [m/s]
Marsta 29 m (4 y)							
850 hPa (4 y)	7.21	3.97	2.70	0.51	0.60	-0.86	1.80
42 m Sigma (4 y)	6.83	3.10	2.91	0.55	0.84	-1.34	1.78
Other Data (4 y)	4.98	1.93	2.84	0.54	0.90	-0.57	2.01

5.2 PRIMARY REGRESSION RESULTS

Notice that only OLS, LAD and RMA are used in the MCP methods later on.

Table 4. Primary regression results for Ringhals measuring mast 96 meters above ground with different methods and 9 years of data used.

	850 hPa (9 y)			42 m Sigma (9 y)			Other Data (9 y)		
Wind data series	Coeff (x^2)	Coeff (x)	Y-intercept	Coeff (x^2)	Coeff (x)	Y-intercept	Coeff (x^2)	Coeff (x)	Y-intercept
OLS		1.006	0.316		0.643	1.100		0.938	0.483
LAD		1.049	0.075		0.638	0.978		0.938	0.404
RMA		0.812	1.194		1.240	0.279		0.992	0.067
Curve linear	0.012	0.427	3.312	0.003	0.940	1.980	0.015	0.657	1.600
Bootstrap		1.006	0.317		0.643	1.100		0.938	0.484
Jackknife		1.006	0.316		0.643	1.100		0.938	0.483

Table 5. Primary regression results for Oskarshamn measuring mast 100 meters above ground with different methods and 8 years of data used.

	850 hPa (8 y)			42 m Sigma (8 y)			Other Data (8 y)		
Wind data series	Coeff (x^2)	Coeff (x)	Y-intercept	Coeff (x^2)	Coeff (x)	Y-intercept	Coeff (x^2)	Coeff (x)	Y-intercept
OLS		1.322	-0.689		1.009	-0.048		1.208	-0.381
LAD		1.285	-0.492		1.012	-0.161		1.223	-0.568
RMA		0.555	2.212		0.783	1.478		0.720	1.173
Curve linear	0.001	0.404	3.466	-0.01	0.666	2.457	-0.01	0.710	1.605
Bootstrap		1.323	-0.689		1.009	-0.046		1.208	-0.381
Jackknife		1.322	-0.689		1.009	-0.048		1.208	-0.381

Table 6. Primary regression results for Marsta measuring mast 29 meters above ground with different methods and 4 years of data used.

	850 hPa (4 y)			42 m Sigma (4 y)			Other Data (4 y)		
Wind data series	Coeff (x^2)	Coeff (x)	Y-intercept	Coeff (x^2)	Coeff (x)	Y-intercept	Coeff (x^2)	Coeff (x)	Y-intercept
OLS		1.196	1.994		1.306	1.122		0.867	1.190
LAD		1.227	1.794		1.347	0.888		0.871	1.125
RMA		0.503	0.738		0.646	-0.043		1.037	-0.794
Curve linear	0.009	0.157	2.634	0.010	0.398	1.115	0.018	0.736	0.197
Bootstrap		1.196	1.996		1.305	1.121		0.867	1.192
Jackknife		1.196	1.994		1.306	1.122		0.867	1.190

5.3 SECONDARY REGRESSION RESULTS

If there is no deviation between the measured and the estimated mean wind speed, the value is set to 100.00; if there is an underestimation the value will be lower than 100 and for the case where an overestimation occurs the value is higher than 100. The explanation behind the Weibull distribution can be obtained from section 3.2 Wind Distribution, and the explanation behind the used regression methods can be found in section 3.5.3 Linear and Curve Estimations.

Table 7. Secondary regression results for Ringhals measuring mast at 96 meters above ground.

	Normalized values [%]				Non-normalized values [m/s]	
Wind speed series after MCP	$\frac{\bar{X}}{\hat{\bar{X}}}100$	$\frac{\bar{\sigma}}{\hat{\bar{\sigma}}}100$	$\frac{\bar{k}}{\hat{\bar{k}}}100$	$\frac{\bar{c}}{\hat{\bar{c}}}100$	$\sigma(k)$	$\sigma(c)$
Ringhals 96 m (9 y)						
OLS						
850 hPa (30 y)	100.40	124.47	99.93	77.99	0.25	0.25
42 m Sigma (30 y)	101.74	130.58	101.62	76.01	0.32	0.12
Other Data (20 y)	102.13	111.00	102.17	91.82	0.51	0.22
LAD						
850 hPa (30 y)	100.89	119.46	100.77	83.04	0.24	0.25
42 m Sigma (30 y)	104.78	131.39	104.90	78.57	0.32	0.12
Other Data (20 y)	103.21	111.03	103.26	92.89	0.51	0.22
RMA						
850 hPa (30 y)	100.26	101.83	100.25	98.76	0.20	0.27
42 m Sigma (30 y)	101.38	103.65	101.40	97.96	0.25	0.13
Other Data (20 y)	101.98	103.37	101.96	99.11	0.47	0.23

Table 8. Secondary regression results for Oskarshamn measuring mast at 100 meters above ground.

	Normalized values [%]				Non-normalized values [m/s]	
Wind speed series after MCP	$\frac{\bar{X}}{\hat{\bar{X}}}100$	$\frac{\bar{\sigma}}{\hat{\bar{\sigma}}}100$	$\frac{\bar{k}}{\hat{\bar{k}}}100$	$\frac{\bar{c}}{\hat{\bar{c}}}100$	$\sigma(k)$	$\sigma(c)$
Oskarshamn 100 m (8 y)						
OLS						
850 hPa (30 y)	98.73	135.38	99.83	71.87	0.22	0.33
42 m Sigma (30 y)	96.23	125.07	97.29	76.14	0.19	0.22
Other Data (20 y)	101.32	118.11	102.16	85.39	0.37	0.19
LAD						
850 hPa (30 y)	99.30	139.23	100.38	70.08	0.22	0.32
42 m Sigma (30 y)	97.62	124.77	98.67	77.50	0.18	0.22
Other Data (20 y)	102.34	116.70	103.10	87.31	0.37	0.20
RMA						
850 hPa (30 y)	99.06	99.35	99.22	98.53	0.16	0.39
42 m Sigma (30 y)	97.02	98.86	97.25	96.97	0.14	0.25
Other Data (20 y)	101.15	102.66	101.37	97.86	0.32	0.21

Table 9. Secondary regression results for Marsta measuring mast at 29 meters above ground.

	Normalized values [%]				Non-normalized values [m/s]	
Wind speed series after MCP	$\frac{\bar{X}}{\hat{\bar{X}}}100$	$\frac{\bar{\sigma}}{\hat{\bar{\sigma}}}100$	$\frac{\bar{k}}{\hat{\bar{k}}}100$	$\frac{\bar{c}}{\hat{\bar{c}}}100$	$\sigma(k)$	$\sigma(c)$
Marsta 29 m (4 y)						
OLS						
850 hPa (30 y)	109.63	154.09	107.55	66.86	0.16	0.11
42 m Sigma (30 y)	94.40	111.92	94.39	83.66	0.16	1.51
Other Data (20y)	98.82	109.03	98.92	91.19	0.28	0.23
LAD						
850 hPa (30 y)	109.39	151.39	107.58	68.25	0.16	0.11
42 m Sigma (30 y)	95.31	108.87	95.49	87.66	0.16	1.56
Other Data (20y)	100.00	108.52	100.31	93.07	0.28	0.23
RMA						
850 hPa (30 y)	101.25	97.83	101.65	106.36	0.09	0.16
42 m Sigma (30 y)	95.05	94.83	95.46	102.48	0.14	1.75
Other Data (20y)	98.77	98.38	99.11	102.61	0.25	0.24

5.4 RESIDUAL EVALUATION

The residuals contain important information about how correctly a method describes the reality. In the ideal case the residuals have a high kurtosis value, and the skewness value, mean and median are close to zero. It should be noted that the extent of the residuals in some way cannot only describe the method properties, but also be sensitive to outliers to a high degree.

Table 10. Residual evaluation of regression methods used on wind data from Ringhals measuring mast 96 meters above ground.

Wind speed series after MCP	Kurtosis of residuals [-]	Skewness of residuals [-]	Mean of residuals [m/s]	Median of residuals [m/s]	Highest residual [m/s]	Lowest residual [m/s]
Ringhals 96 m						
OLS						
850 hPa	3.70	0.11	0.00	-0.07	10.42	-9.68
42 m Sigma	3.95	0.60	0.00	-0.16	9.01	-5.62
Other Data	4.71	0.50	0.00	-0.08	8.48	-4.88
LAD						
850 hPa	3.78	0.09	0.05	0.00	10.59	-9.76
42 m Sigma	3.95	0.60	0.15	0.00	9.18	-5.45
Other Data	4.71	0.50	0.08	0.00	8.56	-4.80
RMA						
850 hPa	3.97	-0.03	0.00	0.01	11.08	-11.12
42 m Sigma	3.54	0.32	0.00	-0.08	8.36	-6.21
Other Data	4.71	0.47	0.00	-0.09	8.82	-5.13

Table 11. Residual evaluation of methods used on wind data from Oskarshamn measuring mast 100 meters above ground.

Wind speed series after MCP	Kurtosis of residuals [-]	Skewness of residuals [-]	Mean of residuals [m/s]	Median of residuals [m/s]	Highest residual [m/s]	Lowest residual [m/s]
Oskarshamn 100 m						
OLS						
850 hPa	3.44	0.04	0.00	-0.02	14.76	-11.37
42 m Sigma	3.38	0.21	0.00	-0.10	10.19	-6.10
Other Data	3.42	0.34	0.00	-0.08	6.97	-5.09
LAD						
850 hPa	3.43	0.07	0.05	0.00	14.66	-11.19
42 m Sigma	3.38	0.20	0.10	0.00	10.29	-6.02
Other Data	3.44	0.32	0.09	0.00	7.11	-5.03
RMA						
850 hPa	3.63	-0.27	0.00	0.12	16.78	-12.98
42 m Sigma	3.52	0.01	0.00	0.01	11.32	-7.15
Other Data	3.60	0.20	0.00	-0.03	7.61	-5.94

Table 12. Residual evaluation of methods used on wind data from Marsta measuring mast 29 meters above ground.

Wind speed series after MCP	Kurtosis of residuals [-]	Skewness of residuals [-]	Mean of residuals [m/s]	Median of residuals [m/s]	Highest residual [m/s]	Lowest residual [m/s]
Marsta 29 m						
OLS						
850 hPa	2.65	0.17	0.00	-0.05	11.11	-8.75
42 m Sigma	4.41	0.25	0.00	-0.06	8.19	-6.91
Other Data	5.44	0.51	0.00	-0.05	5.39	-3.64
LAD						
850 hPa	2.66	0.15	0.06	0.00	11.16	-8.75
42 m Sigma	4.47	0.24	0.06	0.00	8.26	-6.94
Other Data	5.45	0.50	0.05	0.00	5.44	-3.60
RMA						
850 hPa	3.10	-0.17	0.00	0.11	10.93	-11.22
42 m Sigma	4.58	0.15	0.00	0.01	8.26	-7.45
Other Data	5.41	0.42	0.00	-0.04	5.46	-3.86

5.5 EMPIRICAL RESULTS

The empirical results are divided into two parts. The first part describes how extending the measurement series affects the deviation from the measured mean wind speed both in graphical form and in table form for some important key months. This makes it easier to relate to other measurement studies.

The second part describes how the use of only one month of the year will affect the deviation from the measured mean wind speed. This means that all the January month data have been selected with their corresponding reference series, and a MCP has been performed. The deviation is not corrected for negative values in the same way as in part one, which makes it possible to see if there is an overestimation (negative percentage values) or an underestimation (positive percentage values). See Equation 20, Equation 21 and Equation 22 for clarification and comparison of the different calculation methods used in these two parts.

5.5.1 Length of the measurement series

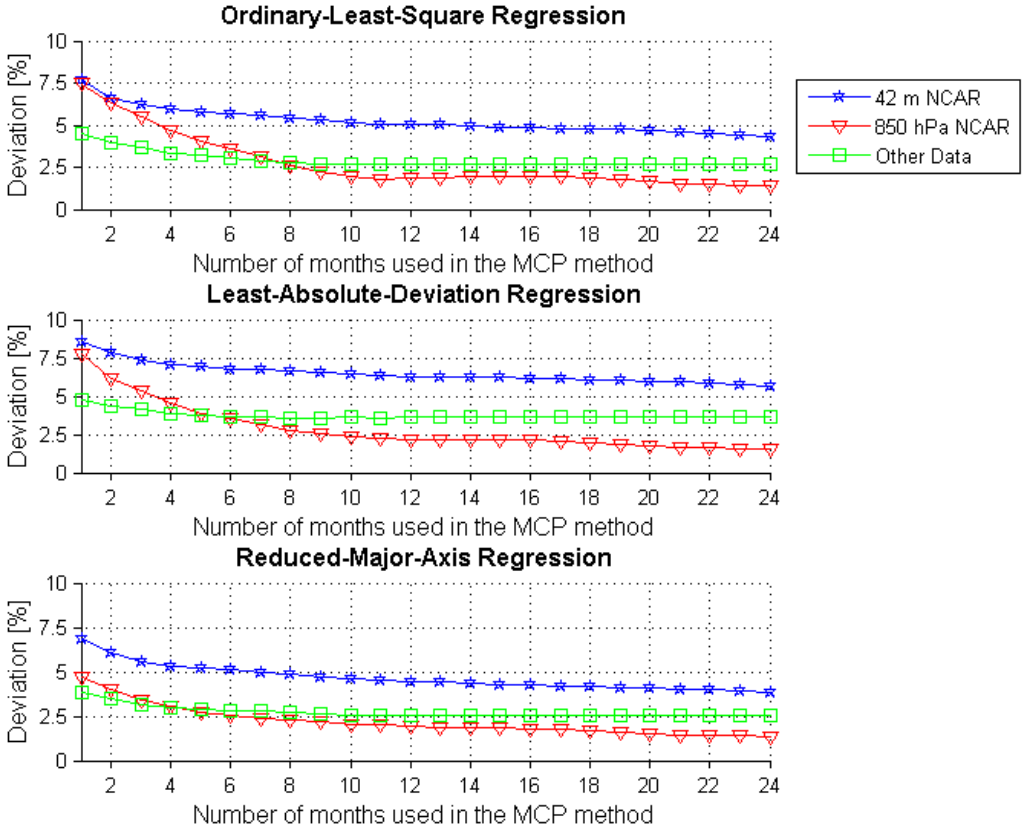

Figure 5. Deviation in percentage from the measured mean wind speed at Ringhals at height 96 meters above ground when increasing the number of months used as input into the MCP methods.

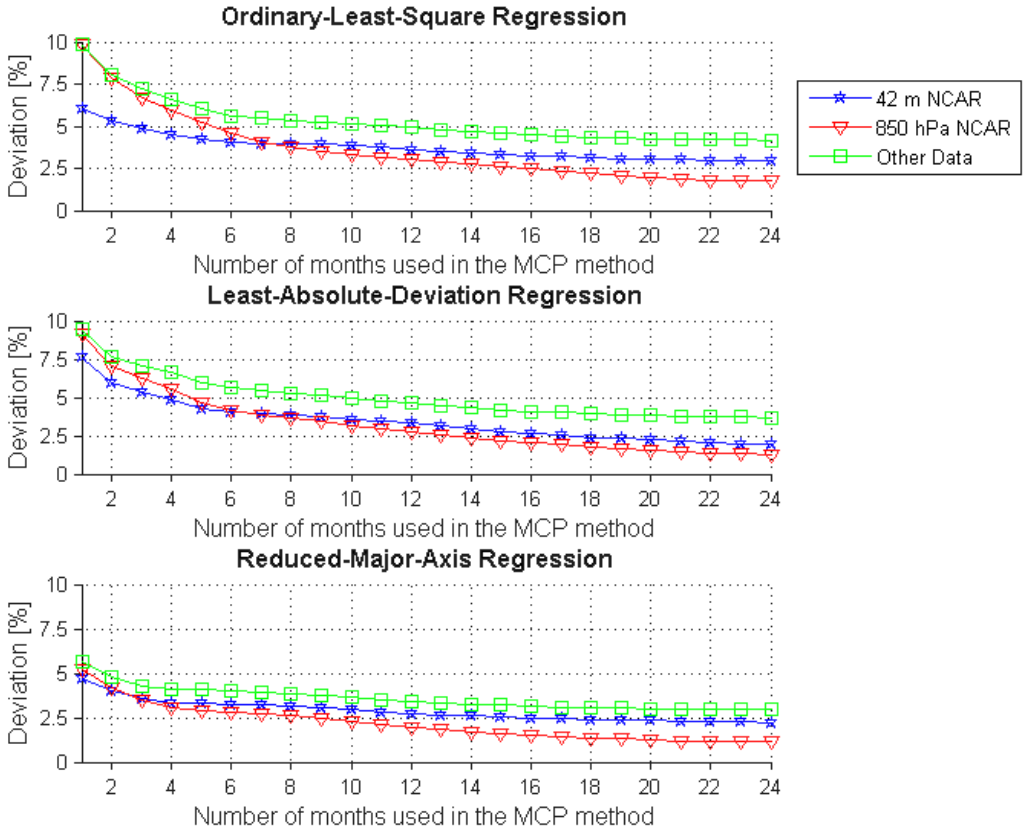

Figure 6. Deviation in percentage from the measured mean wind speed at Oskarshamn at height 100 meters above ground when increasing the number of months used as input into the MCP methods.

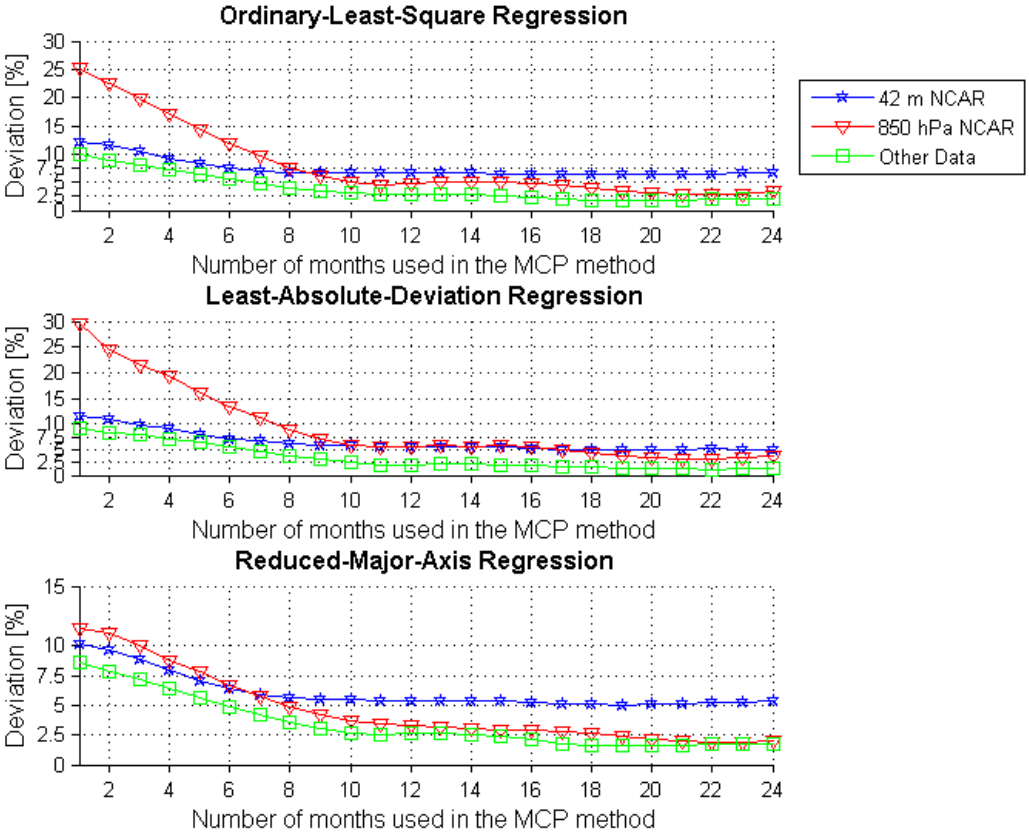

Figure 7. Deviation in percentage from the measured mean wind speed at Marsta at height 29 meters above ground when increasing the number of months used as input into the MCP methods.

5.5.2 Seasonal dependence of the measurement series

The regression methods are applied on the different reference series. The calculations are based on the same model as in section 4.5.1 but here the input into the MCP has been e.g. all Januaries with their corresponding reference series. This returns a long-term corrected wind speed series based on the behavior of this month. The mean value of this long-term corrected wind speed series has been used for calculations according to Equation 21. This equation returns the deviation from the measured mean value of the mast. Negative deviations correspond to an overestimation and positive deviations correspond to underestimation; see Equation 21.

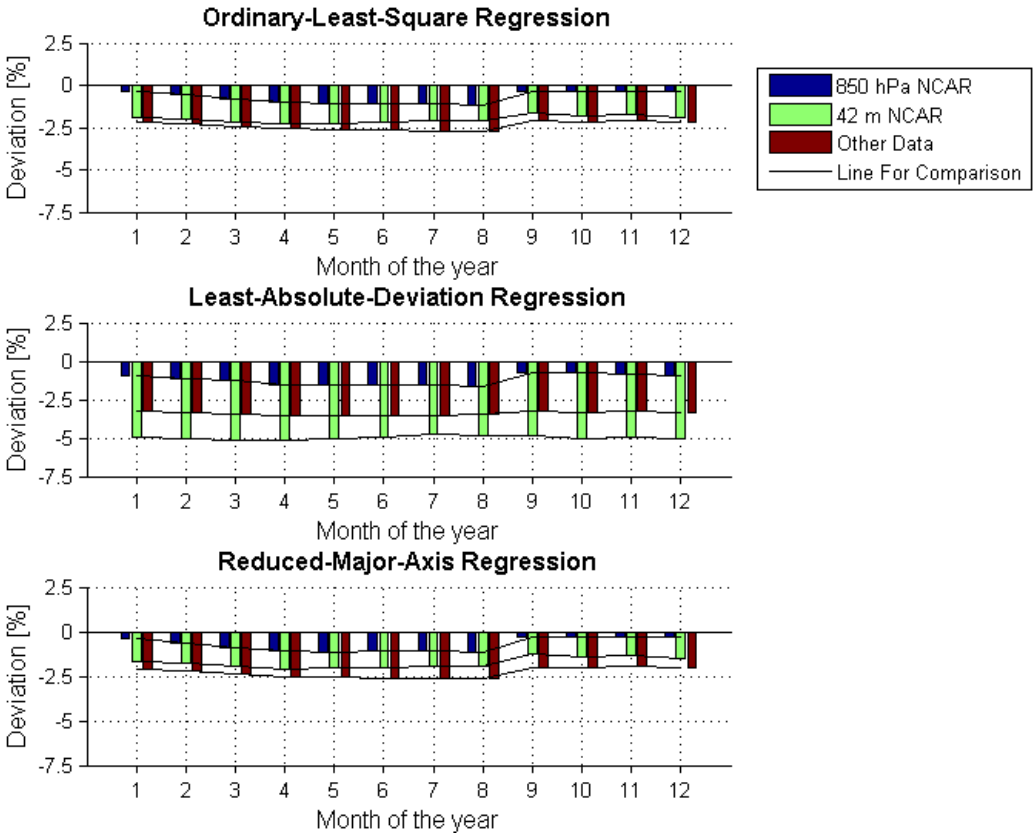

Figure 8. The deviation in percentage from the measured mean wind speed at Ringhals 96 m above ground related to different regression methods when only using one month at a time.

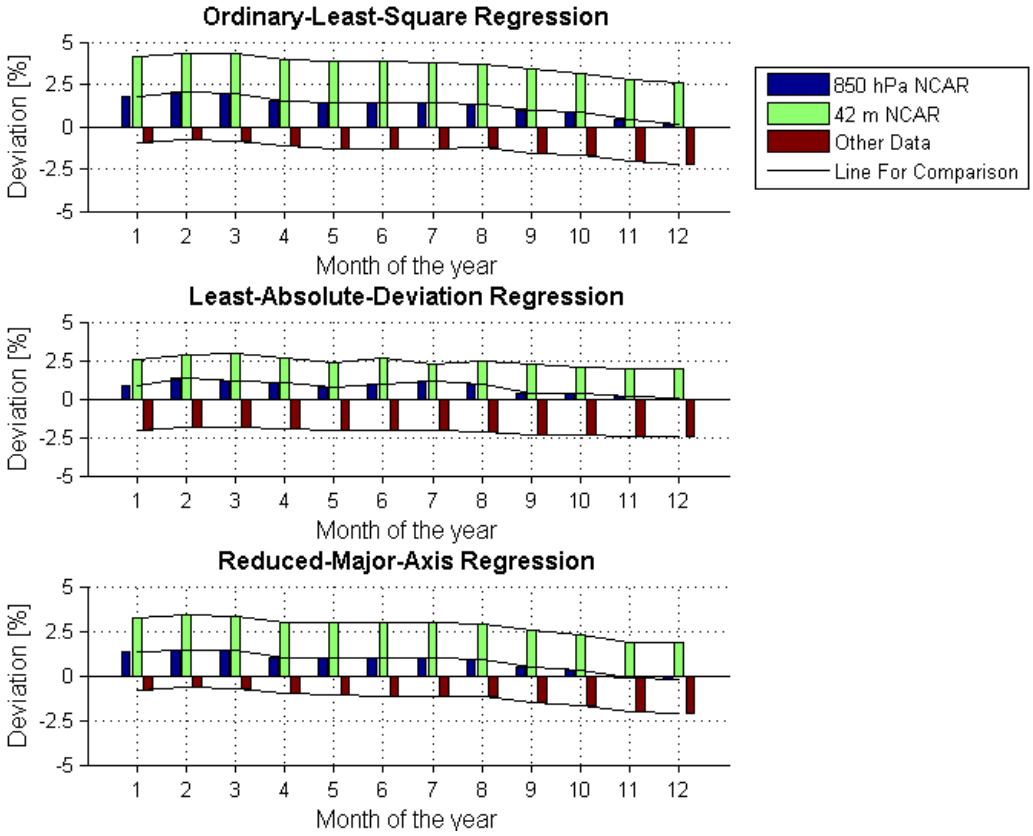

Figure 9. The deviation in percentage from the measured mean wind speed at Oskarshamn 100 m above ground related to different regression methods when only using one month at a time.

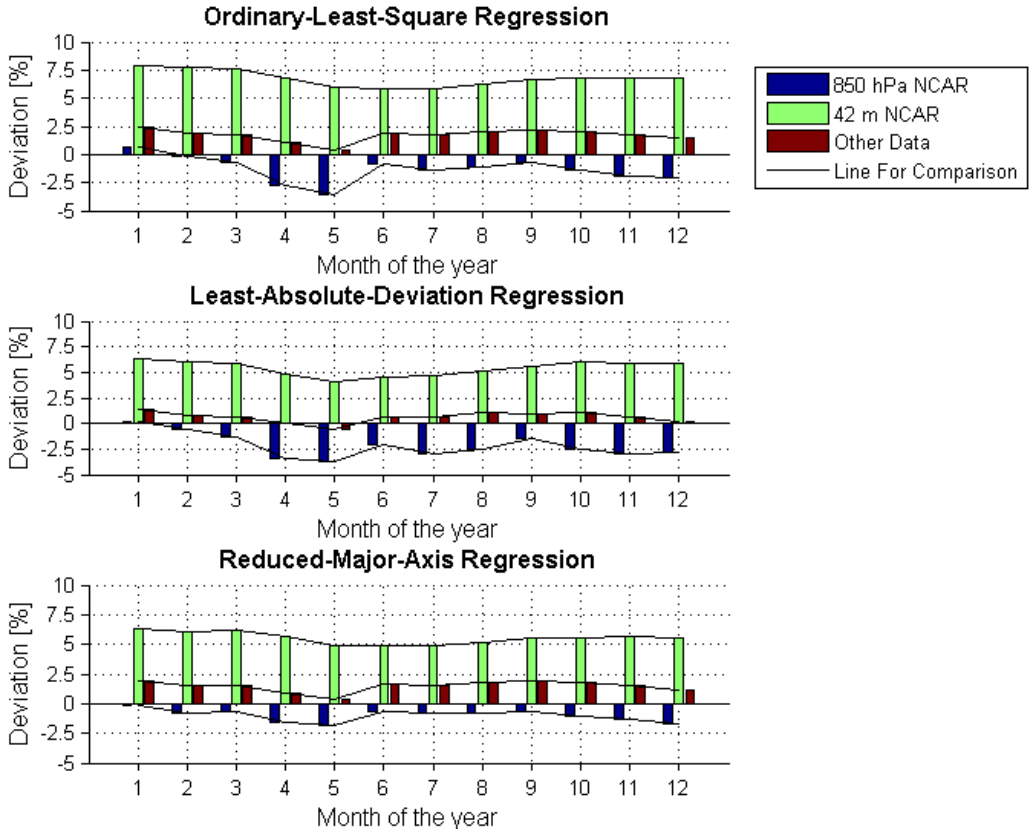

Figure 10. The deviation in percentage from the measured mean wind speed at Marsa 29 m above ground related to different regression methods when only using one month at a time.

6 DISCUSSION

Parts of the conclusions stated here have been removed in the official version due to confidentiality.

6.1 SOURCES OF ERROR

The data available were treated with a method called Cook's distance, which detects outliers. When using methods that detect outliers there is always a risk of removing not only outliers but also non-outliers. But due to the small number of points removed, this is considered to be a small source of error.

The model does not consider the wind direction or the temperature influence. There has only been a manual control of these variables to state that the wind direction is roughly the same. This can be a source of falsely used values, e.g. using a value from a reference series when the wind direction is not the same as the measured wind direction, but this influence is estimated to be small.

The data have been transformed from their original time series into daily mean values of wind speed and wind direction. The data series could be transformed into 6-hour data, which is used in the NCAR data within this model, or even smaller time intervals like 5 to 10-minute values.

Notice should be taken to the length of the measurement series, 9 years for Ringhals, 8 years for Oskarshamn and only 4 years for Marsta.

6.2 REFERENCE SERIES

In Tables 1, 2 and 3 the basic statistic parameters can be studied for the measuring masts Ringhals, Oskarshamn and Marsta. The mean wind speed shown in the first column is expected to be highest for the NCAR 850 hPa data because of being situated much higher into the atmosphere than the measuring mast. The NCAR 42-meter sigma level data are expected to be lower than the Ringhals and the Oskarshamn measuring masts, but higher than the Marsta measuring mast. This was true for Ringhals and Marsta but not for Oskarshamn where the NCAR 42-meter sigma level data were closest to the measured mean wind speed followed by the Other Meteorological Data and the NCAR 850 hPa data. This indicates that all three data series overestimate the mean wind speed at this position.

The standard deviation is expected to follow the same pattern as for the mean wind speed, namely the NCAR 850 hPa data with the highest standard deviation and the Other Meteorological Data with the lowest. This was correct for the measuring mast at Marsta, but for the other two, the NCAR 42-meter sigma level data had the lowest standard deviation followed by the Other Meteorological Data and the NCAR 850 hPa data.

In terms of kurtosis and skewness none of the reference series were able to describe these at the measuring masts in Oskarshamn or Marsta. The NCAR 42-meter sigma level data were able to describe roughly both the kurtosis and the skewness at the Ringhals measuring mast. See Figures 1 and 2 for the relation between kurtosis and

skewness regarding the mean wind speed and the scale and shape factors of the Weibull distribution.

The correlation coefficient is generally highest for the Other Meteorological Data followed by the NCAR 42-meter sigma level data and last the NCAR 850 hPa data.

The last column shows the cross validation results for the respective reference series. It seems that the NCAR 850 hPa data and the Other Meteorological Data follow each other and that the NCAR 42-meter sigma level data contain a larger error at Ringhals and Marsta measuring masts and almost no error at the Oskarshamn measuring mast.

6.3 REGRESSION METHODS

When studying the regression equations in Tables 4, 5 and 6 one sees that the OLS and the LAD regression equations follow each other with some variation in the y-intercept and the slope. The RMA has significantly different values. It is also stated that there is a weak relationship to the curve linear equation. The two resample methods used, the Bootstrap and the Jackknife, follow the OLS almost exactly in all cases.

The next step of the regression analysis is to study the results after a MCP has been used. This is shown in Tables 7, 8 and 9. Common for all reference series is that when using the RMA regression method, the overall deviation decreases compared to both the OLS and the LAD. The RMA regression method is able to estimate the mean wind speed with a relatively low standard deviation and both Weibull parameters with good accuracy. It is interesting to see that all reference series in combination with the OLS and the LAD regression methods fail to estimate the Weibull shape factor, but all the reference series were able to estimate the Weibull scale factor using the RMA regression method. A maximum deviation of 96. 97 % when excluding the NCAR 850 hPa data from the Marsta measuring mast (106. 36 %) could be seen. Exclusion of the NCAR 850 hPa data should be done because of the obvious poor representation of this reference series. This means that the same reference series can describe the measured values almost accurately with one method and not at all with another.

When comparing the reference series with each other it was clear that the NCAR 850 hPa data had the lowest deviation with the Ringhals measuring mast and the Oskarshamn measuring mast data.

Regarding the Marsta measuring mast the Other Meteorological Data were near the measured value with all methods, and the NCAR 850 hPa data were always far from the measured value. But when using the RMA method there was only 0. 02 % difference in deviation between the NCAR 850 hPa data (101. 25 %) and the Other Meteorological Data (98. 77%) if 100. 00 % was considered.

When studying the Ringhals measuring mast all methods and reference series seem to overestimate the mean wind speed. For the Oskarshamn and Marsta measuring masts the NCAR 42-meter sigma level data underestimates the mean wind speed. The NCAR 850 hPa data underestimates the mean wind speed at the Oskarshamn measuring mast and overestimates it at the Marsta measuring mast. The Other Meteorological Data

behave the opposite way, overestimating the mean wind speed at the Oskarshamn measuring mast and underestimating it at the Marsta measuring mast.

To validate the used methods a residual validation has been compiled, which can be studied in Table 10, Table 11 and Table 12. The general pattern of the residuals reveals that the Other Meteorological Data have the highest kurtosis values, the NCAR 850 hPa has the lowest skewness values, the LAD regression method has the median equal to zero and the OLS and the RMA regression methods have the mean equal to zero, the NCAR 850 hPa data have the widest distance between the residuals but are also the most symmetric and the Other Meteorological Data have the closest distance between the residuals. The residuals show how well a method expresses the measured data. With a high kurtosis value the residuals have a higher concentration around their mean value. By also having the closest distance between the residuals, the Other Meteorological Data have high prediction ability. The skewness values of the Other Meteorological Data are the highest, and the NCAR 850 hPa data have the lowest values. This indicates that even though the Other Meteorological Data have a high concentration of residuals around the mean, the NCAR 850 hPa data have more evenly distributed residuals around the mean value. It seems that the Other Meteorological Data have higher concentration of the residuals around their mean value, but in some way overestimate the mean wind speed, and the NCAR 850 hPa data have less centered residuals and less over or underestimation.

6.4 LENGTH OF THE MEASUREMENT SERIES

When studying the deviation from the measured mean wind speed with regards to the measuring length for the Ringhals measuring mast, the Other Meteorological Data give the lowest initial deviation, but after four to eight months the NCAR 850 hPa data have a lower deviation. The RMA regression method gives the lowest initial deviation regardless of the reference series.

Regarding the Oskarshamn measuring mast the results are the same for the methods as in the Ringhals case; the RMA regression method gives the lowest deviation regardless of the reference series. The reference series giving the lowest initial deviation is the NCAR 42-meter sigma level. After eight months the NCAR 850 hPa data decline lower than the NCAR 42-meter sigma level data when using the OLS regression method. When using the LAD regression method it takes seven months for the NCAR 850 hPa data to get a lower deviation than the 42-meter sigma level, and for the RMA regression method it only takes three months for the NCAR 850 hPa data to go under the deviation level of the NCAR 42-meter sigma level data.

The measuring mast at Marsta has different results than the first two. The reference series that gives the lowest deviation from the mean wind speed is the Other Meteorological Data, regardless of method used. The RMA regression method gives the lowest initial deviation. There are some fluctuations in deviation between the methods when using the NCAR 42-meter sigma level data and the Other Meteorological Data. When using the NCAR 850 hPA data the RMA regression method shows the lowest deviation regardless of how many months are used. When using the NCAR 42-meter

sigma level data and the Other Meteorological Data, the RMA and the LAD regression methods seem to cross each other more than once. This makes it difficult to draw any certain conclusions in this case.

Regardless of the method and reference series used on all the measuring masts, there is stagnation in the deviation around 12 months of used data. Before and after 12 months there is a continuous decrease in the deviation. This is especially clear in the Marsta case with a drastic plateau from 10 to 14 months of used data. It should be noted that Marsta only had half the length of the other two measurement series.

6.5 SEASONALITY

Figure 14, see Appendix B, shows a seasonal dependence for all of the reference data series at the measuring mast at Ringhals with a greater overestimation at April, June, July and August than during the other months of the year. The period with least deviation was September, October, November and December. This seems to be the pattern for the Ringhals measuring position. A drop in deviation can be seen between August and September, with August having the highest deviation and September the lowest. Starting in September, the deviation seems to increase till August.

The pattern is the same for all reference series and all methods at this position; see Figure 8. This means that a measurement series from September to December would give a less deviated result than a measurement series from May to August, regardless of reference series or method used.

On the site for the Oskarshamn measuring mast there seems to be two drops in deviation for all data sources; see Figure 15, Appendix B. One occurs in February-March and the other in August. When using the Other Meteorological Data it seems to be an inverted behavior, which means an increase in deviation in February-March and in August. The Other Meteorological Data also overestimate the mean wind speed when both the NCAR 850 hPa and the 42-meter sigma level data series underestimate the mean wind speed.

In the positions for Ringhals and Oskarshamn the NCAR 850 hPa data have the lowest deviation with all methods followed by the Other Meteorological Data and the 42-meter sigma level data. When comparing the methods the RMA regression method gave the lowest deviation in all cases, followed by the OLS and the LAD. Conclusively the NCAR 850 hPa data in combination with the RMA regression method gave the least deviations in general, although the Other Meteorological Data in some cases gave nearly as good and within a few months even better results with the OLS and the RMA regression methods.

Studying the Marsta position reveals different results . The data source that reduced the deviation the most was the Other Meteorological Data, which had a low deviation with all methods used. The NCAR 850 hPa data had a low deviation but not as low as the Other Meteorological Data, and the NCAR 850 hPa data had greater fluctuations. The NCAR 42-meter sigma level data in general placed far above the other two data sources; see Figure 16 and Figure 10. In one case the NCAR 42-meter sigma level data had the lowest deviation, when using the RMA regression method.

The highest deviation is noticed in January and the lowest value is in May for the Other Meteorological Data and the NCAR 42-meter sigma level data; this is opposite from what the first two positions concluded. The NCAR 850 hPa data also gave a change in deviation in May, but instead of showing a decrease there is an increase in the deviation.

From the results in section 5.5.2 Seasonal dependence of the measuring series, it can be concluded that how good result a measuring series will give regarding long-term correction depends on the time of the year the measurement is made. Details of this dependence cannot be known without further studies.

7 CONCLUSIONS

The main results in this thesis are stated in the list below.

- The NCAR 850 hPa Reanalysis-2 data series can be used as a trustworthy reference series in wind assessment studies.
- There is a clear stagnation in uncertainty reduction when using 10 to 14 months of data in Measure-Correlate-Predict methods regardless of method and reference series.
- The RMA regression method is the most effective. It can predict the Weibull parameters and the mean wind speed.
- The time of the year the measurement period is measured affects the uncertainty in the mean wind speed; how much and in which direction depend on the reference series and the Measure-Correlate-Predict method used.

8 FURTHER WORK

In this thesis the goal has been to study wind data with a purely statistical approach. In further development of this area it can be of interest to include more data in the form of new measuring masts at different positions to help the model predict better results. Adding the wind direction and temperature dependence as variables could do this. This can enhance the model and give it higher resolution and better prediction.

Instead of using the arithmetic mean values for calculations of this kind it can be of interest to use the Weibull mean values. This would change the whole calculation procedure and might give a better result.

Long measurement series of wind data are scarce which makes it hard to make broader conclusions when modeling the wind. To be able to model a behavior of any kind the model needs to be calibrated. The model will only be as good as the calibration data are. After calibrating the model it is possible to use it with shorter measurement series, which is of interest in the rapidly increasing field of wind assessment.

9 REFERENCES

Ackerman, S. & Knox, J., 2003. *Meteorology - Understanding the Atmosphere*. Thompson Learning Inc.

AWS Truewind, 2006. *The Use Of Reanalysis Data For Climate Adjustments*. AWS Truewind.

AWS Truewind, 2008. *Welcome -- AWS openWind*. [Online] AWS Truewind Available at: www.awsopenwind.org [Accessed 9 November 2009].

Burton, T., Sharpe, D., Jenkins, N. & Bossanyi, E., 2001. *Wind Energy: Handbook.* Wiltshire: Anthony Rowe Ltd.

EMD International, 2009. *Windpro.com*. [Online] EMD Available at: www.windpro.com [Accessed 9 November 2009].

Energimyndigheten, 2009. *www.energimyndigheten.se*. [Online] Available at: http://www.energimyndigheten.se [Accessed 21 September 2009].

Evans, J., 2009. *MEXNC, SNCTOOLS, and the NetCDF Toobox*. [Online] Available at: htt://mexcdf.sourceforge.net [Accessed 24 August 2009].

EWEA, 2008. *Cumulative Wind Energy Installations*. EWEA.

Gentle, J.E., 2009. *Computational Statistics*. New York: Springer Science+Business Media, LLC 2009.

GH WindFarmer, 2009. *GH WindFarmer*. [Online] Gerrad Hassan Available at: www.gerradhassan.com/products/ghwindfarmer/ [Accessed 9 November 2009].

Good, P.I. & Hardin, J.W., 2006. *Common Errors In Statistics (And How To Avoid Them) Second Edition*. New Jersey: John Wiley & Sons, Inc. Publication.

Heiberger, R.M. & Holland, B., 2004. *Statistical Analysis and Data Display - An Intermediate Course with Examples In S-Plus, R, and Sas*. New York: Springer Science Business Media Inc.

Johnson, R.A., 2005. *Miller & Freund's Probability And Statistics For Engineers*. Upper Saddle River: Person Education, Inc.

Kalnay, E; Kanamitsu, M; Kistler, R; Collins, W; Deaven, D; Gandin, L; Iredell, M; Saha, S; White, G; Woollen, J; Zhu, Y; Chelliah, M; Ebisusazaki, W; Higgins, W; Janowiak, J; Mo, K C; Ropelwski, C; Wang, J; Leetmaa, A; Reynolds, R; Jenne, Roy; Joseph, Dennis, 1996. The NCEP/NCAR 40-Year Reanalysis Project. 77(No. 3).

Kistler, Robert; Kalnay, Eugenia; Collins, William; Saha, Suranjana; White, Glenn; Woollen, John; Chelliah, Muthuvel; Ebisuzaki, Wesley; Kanamitsu, Masao; Kousky, Vernon; van den Dool, Huug; Jenne, Roy; Fiorino, Michael, 1999. *The NCEP/NCAR 50-Year Reanalysis*. Bulletin of the American Meteorological Society.

Map from Hitta.se, 2009. *www.hitta.se*. [Online] Hitta.se Available at: http://www.hitta.se/LargeMap.aspx?var=sverige [Accessed 13 January 2009].

Nilsson, E. & Bergström, H., 2009. *Från Mätt Vind Till Vindklimat - Normalårskorrigering*. Stockholm: Elforsk AB.

NOAA, 2005. *NCEP/DOE AMIP-II Reanalysis (Reanalysis-2) Home Page*. [Online] National Oceanic And Atmospheric Administration Available at: http://www.cpc.noaa.gov/products/wesley/reanalysis2/ [Accessed 29 September 2009].

Patel, M.R., 2006. *Wind And Solar Power Systems - Design, Analysis, and Operation*. Boca Raton: CRC Press.

Shao, J. & Tu, D., 1995. *The Jackknife And Bootstrap*. New York: Springer-Verlag.

Siddiqi, A.H., Khan, S. & Rehman, S., 2005. Wind Speed Simulation Using Wavelets. 2(2).

Sinclair, A.J. & Blackwell, G.H., 2002. *Applied Mineral Inventory Estimation*. Cambridge: University Press.

The MathWorks, 2005a. *regstats - Functions help file*. The MathWorks, Inc.

The MathWorks, 2005b. *mean - Functions help file*. The MathWorks, Inc.

The MathWorks, 2005c. *var - Functions help file*. The MathWorks, Inc.

The MathWorks, 2005d. *std - Functions help file*. The MathWorks, Inc.

The MathWorks, 2005e. *cov - Functions help file*. The MathWorks, Inc.

The MathWorks, 2005f. *corrcoef - Functions help file*. The MathWorks, Inc.

The MathWorks, 2005g. *skewness - Functions help file*. The MathWorks, Inc.

The MathWorks, 2005h. *kurtosis - Functions help file*. The MathWorks, Inc.

Thøgersen, L.M., Motta, M., Sørensen, T. & Nielsen, P., Undated. *Measure-Correlate-Predict Methods: Case Studies and Software Implimentation*. Aalborg: EMD International.

Thøgersen, M.L., Nielsen, P. & Sørensen, T., 2007. *An Introduction to the MCP Facilities in WindPRO*. Aalborg Ø: EMD International A/S.

Trauth, M.H., 2006. *MATLAB Recipes for Earth Science Second Edition*. Berlin: Springer Science + Business Media.

Urdan, T.C., 2005. *Statistics In Plain English*. New Jersey: Lawrence Erlbaum Associates, Inc.

WindFarm, 2009. *WindFarm : Wind Energy Software: Windfarm Software for Designing, Visualising and Optimising*. [Online] ReSoft Available at: www.resoft.co.uk/English/index.html [Accessed 9 November 2009].

Wizelius, T., 2007. *Vindkaft - I Teori Och Praktik*. Studentlitteratur.

APPENDIX

A. LENGTH OF THE MEASUREMENT SERIES

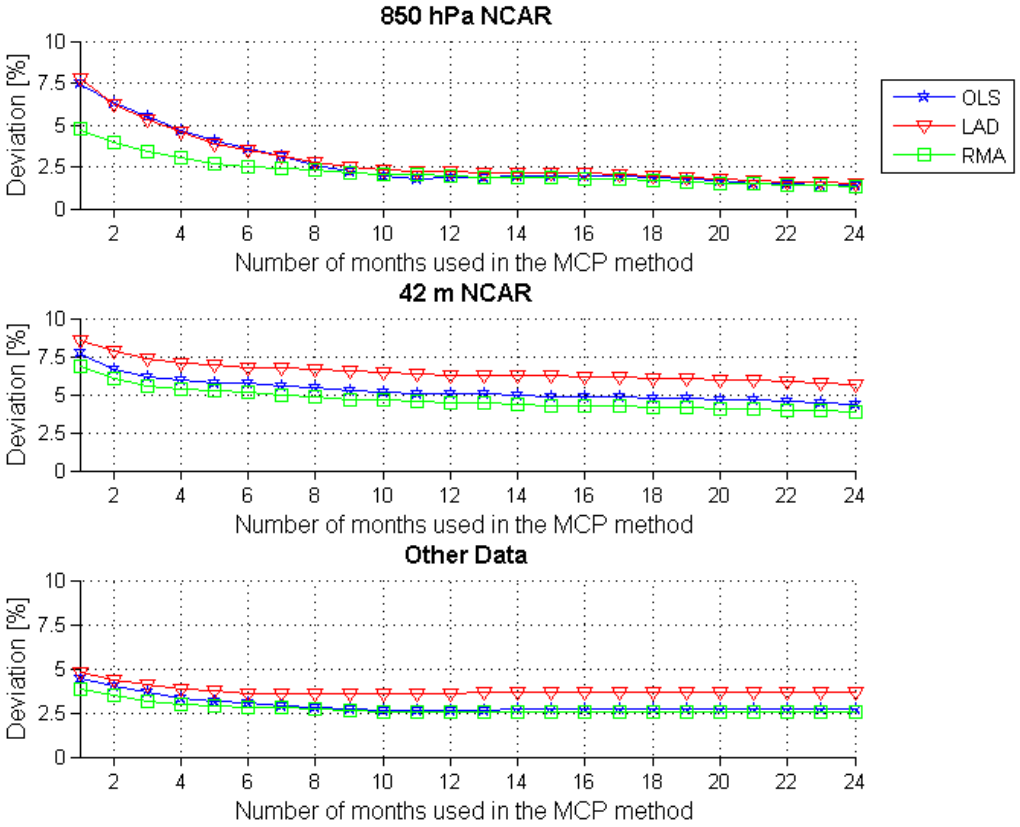

Figure 11. Deviation in percentage from the measured mean wind speed at Ringhals at height 96 meters above ground when increasing the number of months used as input into the MCP methods.

Table 13. Deviation in percentage from measured mean wind speed after extending the number of months measured for Ringhals measuring mast at 96 m above ground. Values are extracted from the curves in Figure 11.

	Deviation in % from measured mean (9 y)								
Number of months	850 hPa NCAR (30 y)			42 m NCAR (30 y)			Other Data (20 y)		
measured	OLS	LAD	RMA	OLS	LAD	RMA	OLS	LAD	RMA
1	7.46	7.79	4.73	7.71	8.59	6.84	4.47	4.78	3.81
3	5.50	5.33	3.42	6.19	7.39	5.54	3.65	4.10	3.17
6	3.60	3.52	2.52	5.71	6.79	5.13	3.03	3.60	2.84
9	2.23	2.52	2.18	5.30	6.52	4.69	2.70	3.58	2.63
12	1.82	2.19	1.93	5.05	6.29	4.45	2.66	3.61	2.50
18	1.88	1.98	1.67	4.79	6.09	4.16	2.69	3.70	2.52
24	1.40	1.50	1.36	4.34	5.67	3.82	2.68	3.65	2.53

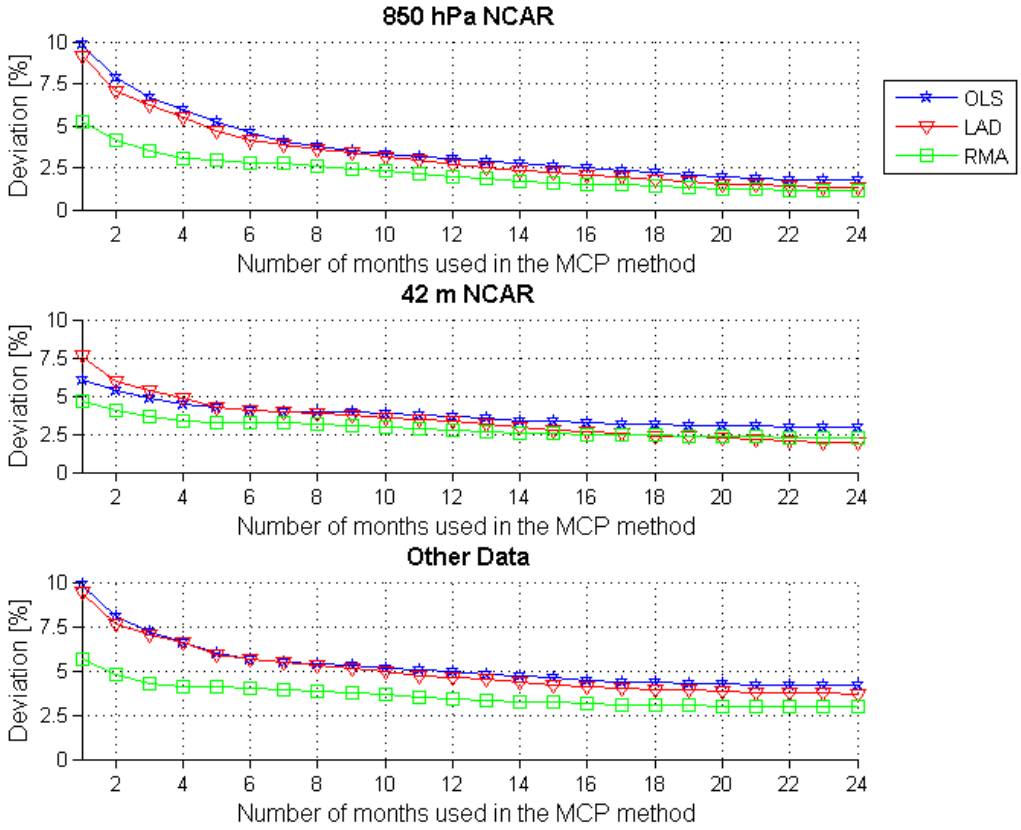

Figure 12 . Deviation in percentage from the measured mean wind speed at Oskarshamn at height 100 meters above ground when increasing the number of months used as input into the MCP methods.

Table 14. Deviation in percentage from measured mean wind speed after extending the number of months measured for Oskarshamn measuring mast at 100 m above ground.

	Deviation in % from measured mean (8 y)								
Number of month	850 hPa NCAR (30 y)			42 m NCAR (30 y)			Other Data (20 y)		
measured	OLS	LAD	RMA	OLS	LAD	RMA	OLS	LAD	RMA
1	9.83	9.15	5.22	6.07	7.62	4.68	9.83	9.41	5.63
3	6.71	6.26	3.53	4.87	5.38	3.62	7.20	7.00	4.27
6	4.58	4.18	2.81	4.08	4.09	3.21	5.61	5.66	4.02
9	3.54	3.44	2.45	3.95	3.73	3.03	5.27	5.11	3.78
12	3.00	2.71	1.96	3.63	3.34	2.74	4.92	4.61	3.41
18	2.22	1.74	1.38	3.12	2.40	2.41	4.33	3.92	3.05
24	1.76	1.29	1.16	2.92	1.92	2.23	4.17	3.69	2.95

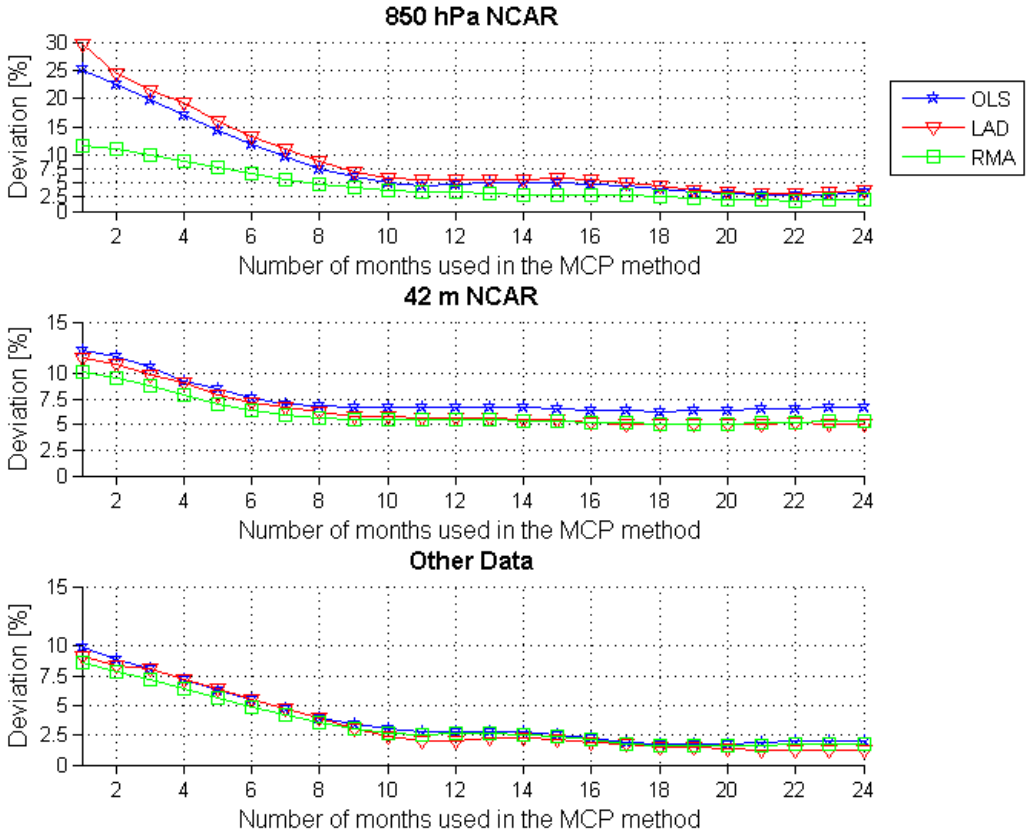

Figure 13. Deviation in percentage from the measured mean wind speed at Marsta at height 29 meters above ground when increasing the number of months used as input into the MCP methods. Here the different reference series are compared with different regression methods.

Table 15. Deviation in percentage from measured mean wind speed after extending the number of months measured for Marsta measuring mast at 29 m above ground.

	Deviation in % from measured mean (4 y)								
Number of months	850 hPa NCAR (30 y)			42 m NCAR (30 y)			Other Data (20 y)		
measured	OLS	LAD	RMA	OLS	LAD	RMA	OLS	LAD	RMA
1	25.12	29.65	11.49	12.22	11.49	10.10	9.90	9.15	8.58
3	19.71	21.37	9.98	10.56	9.83	8.82	8.09	8.03	7.20
6	11.82	13.34	6.65	7.51	7.13	6.35	5.46	5.46	4.89
9	6.07	7.00	4.23	6.70	5.80	5.48	3.44	3.09	3.03
12	4.71	5.54	3.31	6.65	5.61	5.43	2.78	2.03	2.62
18	3.97	4.49	2.60	6.30	5.03	5.05	1.69	1.55	1.62
24	3.38	3.72	2.02	6.65	5.05	5.30	1.95	1.20	1.69

B. SEASONAL DEPENDENCE OF THE MEASUREMENT SERIES

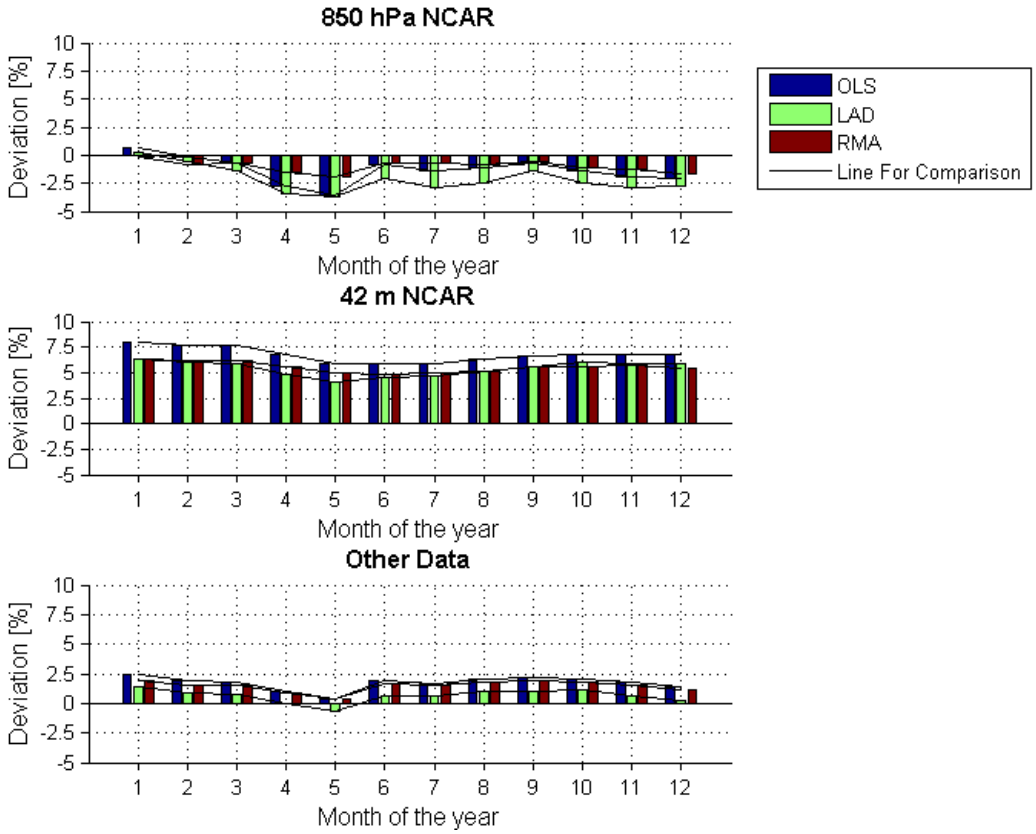

Figure 14. The deviation in percentage from the measured mean wind speed at Ringhals 96 m above ground related to different reference series when only using one month at a time.

Table 16. The deviation in percentage from measured mean wind speed when only using one month at a time for Ringhals 96 meter.

	Deviation in % from measured mean								
	850 hPa NCAR			42 m NCAR			Other Data		
Month of the year	OLS	AMD	RMA	OLS	AMD	RMA	OLS	AMD	RMA
1	-0.39	-0.95	-0.43	-1.92	-4.99	-1.65	-2.19	-3.30	-2.08
2	-0.55	-1.13	-0.66	-1.96	-5.04	-1.80	-2.31	-3.38	-2.23
3	-0.81	-1.26	-0.90	-2.15	-5.19	-1.98	-2.44	-3.43	-2.37
4	-1.02	-1.53	-1.07	-2.30	-5.19	-2.10	-2.59	-3.51	-2.50
5	-1.14	-1.58	-1.14	-2.26	-5.06	-2.03	-2.62	-3.52	-2.53
6	-1.07	-1.55	-1.11	-2.23	-4.95	-2.03	-2.68	-3.56	-2.59
7	-1.10	-1.57	-1.12	-2.09	-4.79	-1.91	-2.69	-3.53	-2.60
8	-1.2	-1.68	-1.19	-2.08	-4.89	-1.91	-2.71	-3.49	-2.61
9	-0.33	-0.79	-0.27	-1.61	-4.88	-1.28	-2.13	-3.28	-1.98
10	-0.34	-0.72	-0.26	-1.81	-5.00	-1.41	-2.17	-3.33	-2.01
11	-0.38	-0.90	-0.28	-1.74	-4.95	-1.35	-2.10	-3.22	-1.94
12	-0.36	-0.92	-0.28	-1.87	-5.09	-1.47	-2.20	-3.31	-2.04

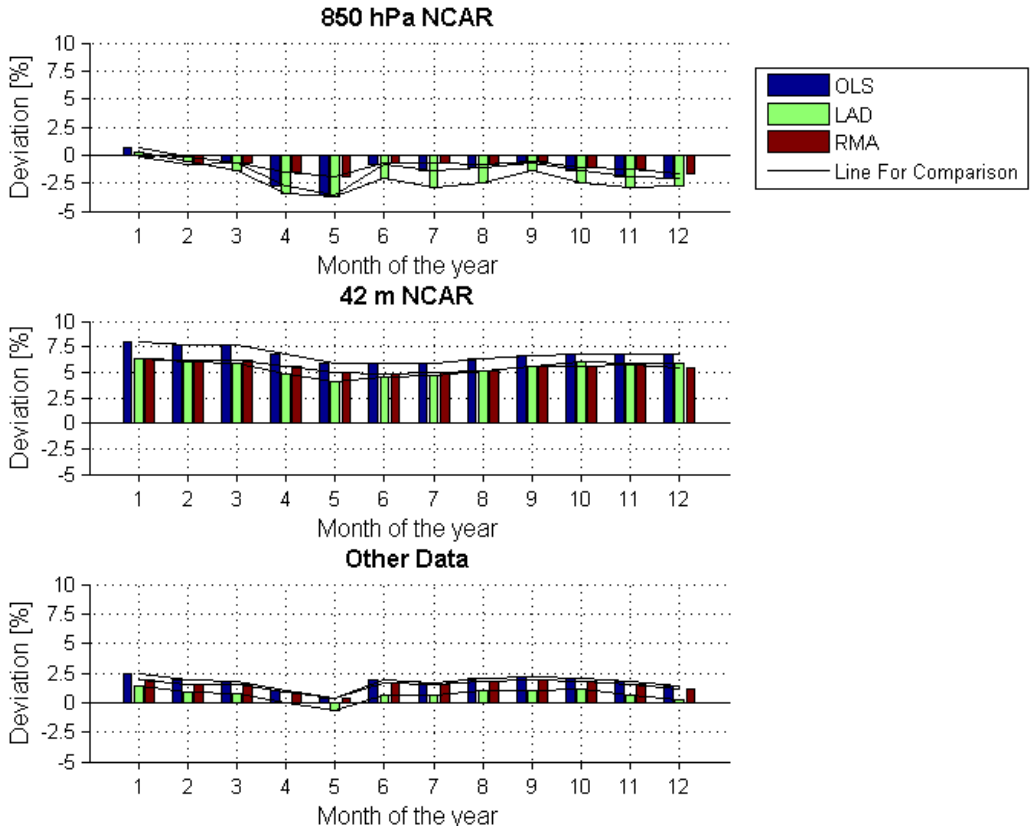

Figure 15. The deviation in percentage from the measured mean wind speed at Oskarshamn 100 m above ground related to different reference series when only using one month at a time.

Table 17. The deviation in percentage from measured mean wind speed when only using one month at a time for Oskarshamn 100 meter.

	Deviation in % from measured mean								
	850 hPa NCAR			42 m NCAR			Other Data		
Month of the year	OLS	AMD	RMA	OLS	AMD	RMA	OLS	AMD	RMA
1	1.74	0.84	1.31	4.12	2.55	3.25	-0.99	-2.02	-0.82
2	2.04	1.38	1.46	4.36	2.87	3.38	-0.76	-1.86	-0.66
3	1.94	1.14	1.39	4.32	2.9748	3.35	-0.83	-1.90	-0.71
4	1.54	1.01	1.03	3.94	2.63	3.01	-1.13	-1.93	-1.02
5	1.37	0.78	0.97	3.84	2.36	2.99	-1.28	-2.08	-1.11
6	1.44	0.98	1.00	3.87	2.61	3.00	-1.28	-2.07	-1.13
7	1.42	1.12	1.01	3.76	2.28	2.93	-1.32	-2.08	-1.15
8	1.33	0.96	0.91	3.72	2.40	2.90	-1.25	-2.11	-1.12
9	0.93	0.39	0.52	3.37	2.21	2.56	-1.58	-2.39	-1.47
10	0.87	0.32	0.32	3.17	2.05	2.28	-1.72	-2.39	-1.66
11	0.43	0.19	-0.12	2.75	1.95	1.88	-2.05	-2.42	-2.02
12	0.18	0.01	-0.23	2.61	1.95	1.84	-2.24	-2.48	-2.15

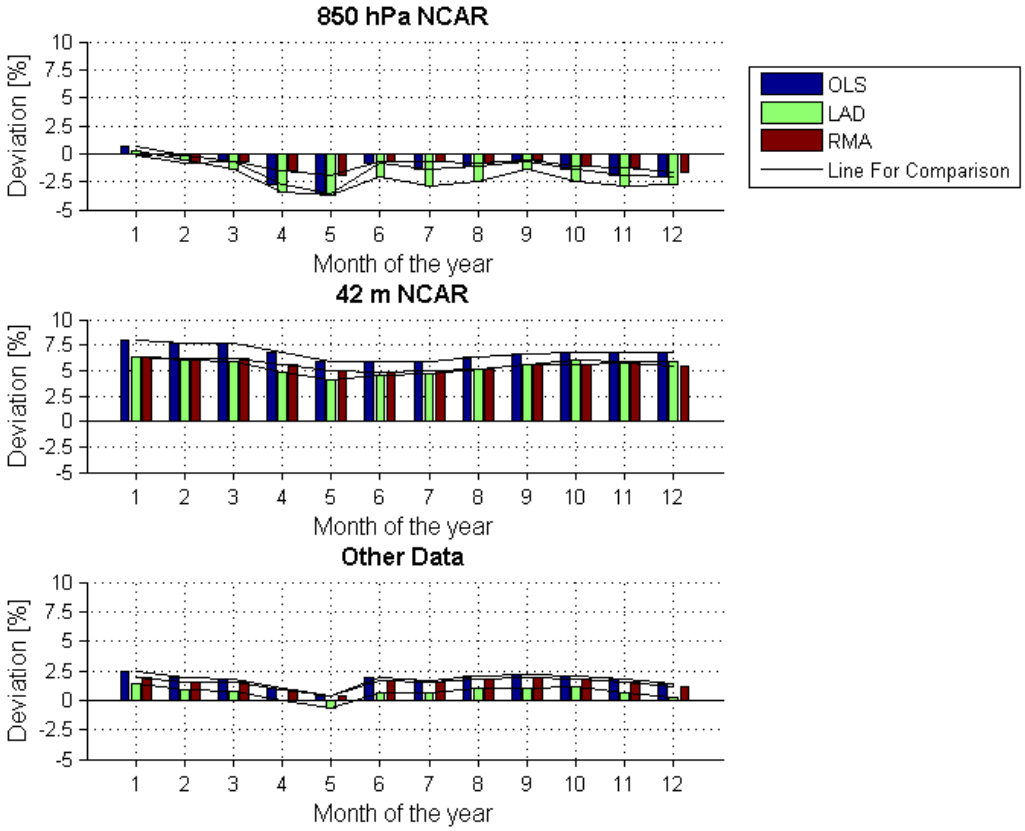

Figure 16. The deviation in percentage from the measured mean wind speed at Marsta 29 m above ground related to different reference series when only using one month at a time.

Table 18. The deviation in percentage from measured mean wind speed when only using one month at a time for Marsta 29 meter.

	Deviation in % from measured mean								
	850 hPa NCAR			42 m NCAR			Other Data		
Month of the year	OLS	AMD	RMA	OLS	AMD	RMA	OLS	AMD	RMA
1	0.69	0.22	-0.20	7.94	6.33	6.30	2.38	1.34	1.93
2	-0.12	-0.60	-0.78	7.73	6.03	6.11	1.90	0.83	1.50
3	-0.64	-1.37	-0.66	7.64	5.91	6.17	1.82	0.70	1.54
4	-2.72	-3.41	-1.57	6.81	4.85	5.62	1.02	0.03	0.91
5	-3.52	-3.73	-1.90	5.94	4.10	4.94	0.35	-0.61	0.35
6	-0.80	-2.08	-0.64	5.85	4.59	4.86	1.88	0.66	1.65
7	-1.32	-2.93	-0.76	5.84	4.73	4.93	1.71	0.64	1.54
8	-1.05	-2.50	-0.84	6.26	5.15	5.19	2.00	1.05	1.74
9	-0.75	-1.43	-0.61	6.66	5.56	5.53	2.14	0.96	1.87
10	-1.34	-2.50	-1.09	6.81	5.95	5.59	2.08	1.11	1.79
11	-1.94	-2.93	-1.29	6.84	5.95	5.66	1.78	0.61	1.56
12	-2.05	-2.77	-1.29	6.78	5.92	5.49	1.43	0.18	1.16