

Short-term Solar Power Forecasting

Peder Bacher

Kongens Lyngby 2008
IMM-2008-13

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

Abstract

The share of the global power production coming from solar power is increasing. Forecasts of solar power is a key point for a successful integration of the solar power production into the existing electricity grid. In the present thesis a solar power forecasting method is developed.

Observations of solar power from 21 grid-connected photovoltaic (PV) systems with peak power in the range of 1 kWp to 4 kWp are used. The PV systems are located within an area of a few square kilometers. Numerical weather predictions (NWP) of global irradiation from the mesoscale NWP model DMI-Hirlam, is also used as input to the method.

A clear sky model only based on solar power observations is developed. It is used to transform the solar power process into a stationary process that resembles the transmittance of the clouds. This process is modeled with linear models and the best model both with and without NWPs as input is identified.

Adaptive estimation is found to be a requisite. Therefore the prediction models of the transformed solar power process are fitted using k-step recursive least squares with forgetting.

The evaluation of the method focus on solar power forecasts for the purpose of bidding into the Nordic electricity market Elspot. The forecasts are issued at 12:00 UTC each day and consist of hourly values of predictions up to a 36 hour horizon of solar power. These forecasts are evaluated and compared to a persistence reference prediction model.

The achieved results clearly indicate an increasing performance for next day horizons (12 to 36 hours), by the model using NWPs as input. Whereas for very short-term predictions (less than 6 hours) the solar power observations are the most important input.

Finally ideas for both refinements and extensions to the method are outlined, together with a suggestion for the development of a framework for standardization of solar power forecasting method evaluation.

-Keywords: *solar power, PV system, photovoltaic, global irradiance, prediction, forecasting, grid-connected, numerical weather prediction, renewable energy.*

Resumé

El-produktion med solceller er på globalt plan i kraftig vækst, men som for andre former for vedvarende energi, som bl.a. vindkraft, er produktionen afhængig af klima variabler med store variationer. Der skal derfor findes metoder til udligning af disse variationer, hvis en fortsat stigende del af el-produktionen skal være vedvarende energi.

Forudsigelser af el-produktionen fra solceller er et vigtigt bidrag til en succesfuld integration af disse i det eksisterende el-net. I dette speciale er udviklet en metode til forudsigelse af el-produktion fra solceller.

Kvarters værdier for et års samlet el-produktion fra 21 solcellesystemer, som er opsat på private hustage i forbindelse med projektet Sol300, er brugt sammen med forudsigelser (numerical weather predictions, NWP) af global indstråling lavet med NWP modellen DMI-Hirlam.

En model, der til et givet tidspunkt estimerer solcelle el-produktionen ved skyfri himmel, kaldes en clear sky model. En clear sky model, som kun er baseret på solcelle el-produktion observationer, er blevet udviklet i specialet. Den bruges til at transformere el-produktion til en stationær process, som derefter modeleres med lineære adaptive modeller.

Fokus er lagt på forudsigelser til brug ved salg af el-produktion på det nordiske el-marked Elspot. Forudsigelserne laves kl. 12:00 UTC hver dag og består af timeværdier for solcelle el-produktionen op til 36 timer frem i tiden.

Forudsigelserne evalueres og sammenlignes med en referencemodel, som er vedholdende el-produktion.

De opnåede resultater indikerer tydeligt, at forudsigelser for næste dag horisonter (12 - 36 timer) bliver væsentligt forbedret ved brug af NWPs som input. For korte horisonter er observationer af solcelle el-produktionen indenfor de forrige par timer det vigtigste input.

Rapporten afsluttes med ideer til forbedring og udvidelse af metoden, sammen med et foreslag om at lave en standard for evaluering af metoder til forudsigelse af solcelle el-produktion.

Preface

This thesis was prepared at Department of Informatics and Mathematical Modeling at the Technical University of Denmark in partial fulfillment of the requirements for acquiring the Master degree in engineering.

I would like to thank my supervisor Henrik Madsen for friendly and informative guidance, Henrik Aalborg Nielsen from IMM and Søren Poulsen from Danish Technological Institute for starting up the project with great energy, and among others at Energimidt, Jacob Vestersager Engdal and Carl Stephansen for being very kind and supportive.

With hopes of a bright future for solar energy and propagation of peace to the entire planet, I thank everybody who have supported me.

Peder Bacher, February 2008.

Contents

Abstract	i
Resumé	iii
Preface	v
Abbreviations	ix
1 Introduction	1
1.1 Motivation	1
1.2 Aim of the thesis	3
1.3 Previous studies of prediction methods for solar power	3
1.4 Thesis outline	4
1.5 Notation	5
2 Global irradiance prediction methods	9
2.1 Parameterized atmosphere simulation and ARIMA(p, d, q) models	9
2.2 Identification of ARMA(p, q) model by the use of a Kalman filter	10
2.3 Forecasting using artificial intelligence techniques	10
2.4 Prediction using sky images	12
2.5 Using satellite images and NWPs	12
3 Data	15
3.1 Scaling and synchronization of data	16
3.2 Sol300	18
3.3 DMI	21
3.4 Power production and irradiance observations	25
4 Clear sky model	27

4.1	Surface estimation	32
4.2	Kernel method	32
4.3	Spline method	33
4.4	Results	35
4.5	Comparison of the estimates	42
4.6	Conclusion	43
5	Modeling procedure	45
5.1	Power forecasts	46
5.2	Model evaluation	47
5.3	Reference model	49
5.4	k -step recursive least squares with forgetting	52
5.5	Finding the best α_{cut} limit	53
5.6	Modeling procedure	55
6	Autoregressive models	57
6.1	Model identification	57
6.2	Evaluation of the selected AR model	62
7	Autoregressive models with exogenous input	67
7.1	Transforming global irradiance to solar power	67
7.2	Identification of the best ARX model	73
7.3	Evaluation of the selected ARX model	76
8	Summary and discussion	81
8.1	Final solar power forecast evaluation	81
8.2	Performance summary numbers	84
8.3	Uncertainty modeling	86
8.4	Solar power forecasts for Elspot	86
8.5	Discussion and further work	89
9	Conclusion	93

Abbreviations

ACF	Autocorrelation Function
AR	Autoregressive
ARIMA	Integrated Autoregressive Moving Average
ARMA	Autoregressive Moving Average
ARX	Autoregressive with exogenous input
CEST	Central European Summer Time
CET	Central European Time
D	Dimension
DMI	Danish Metheological Institute
ECMWF	European Centre for Medium-Range Weather Forecasts
GSM	Global System for Mobile communications
LS	Least Squares
LST	Local Solar Time
NWP	Numerical Weather Prediction
MOS	Model Output Statistics
RLS	Recursive Least Squares
RMSE	Root Mean Square Error
PV	Photovoltaic
TOD	Time Of Day
UTC	Coordinated Universal Time

CHAPTER 1

Introduction

The increasing focus on renewable energy sources motivates studies concerning the integration of these into existing energy systems. Solar energy is along with wind energy expected to play a major role in the future energy supply. The fluctuating nature of the energy output from such sources require reliable forecast information for a successful integration in the electricity grid. The potential of solar energy is very well described by Figure 1.1. In the present thesis a method for forecasting the power output of photovoltaic (PV) systems is developed. The forecasts are of hourly values in the horizons 1 to 36 hours, and they are issued at 12:00 Coordinated Universal Time (UTC) every day. The data used is observations of power production from 21 PV systems (solar power) located within an area of a few square kilometers, and numerical weather predictions (NWP) from the mesoscale NWP model DMI-Hirlam.

1.1 Motivation

The main problem of integrating solar power into the existing electricity grid, is to balance the power put into the grid from other sources. The expenses of changing the output of fossil fuel power plants is increasing as the horizon of the change is decreasing. Thus improved forecasting of solar power can optimize the total power production and forecasting is therefore a requisite to increase the share of solar power put into the electricity grid. In the future when energy

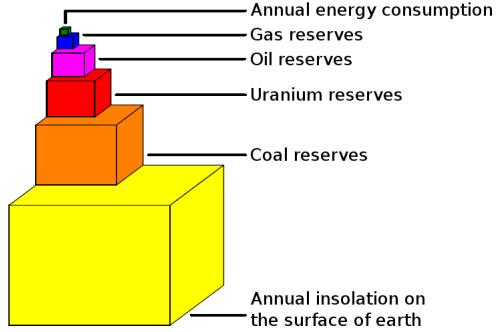


Figure 1.1: *Coarse overview of the proportions between different potential energy resources. It is noted that there is a high uncertainty and many different reports of the amount left of the different resources. The figure is taken from [12]. A discussion of energy reserves and global consumption is found at [25].*

storage technologies are implemented, solar power forecasting will still be an indispensable tool.

The task of balancing the power production in Denmark is done by the Nordic energy market Nord Pool [19]. The electric power is traded on two markets with different horizons. On the Elspot market, hourly power contracts are traded daily before 12:00 Central European Time (CET) for physical delivery the next day. The Elbas market provides continuous power trading 24 hours a day, 7 days a week covering individual hours, up to one hour before delivery. The idea behind the two markets, is that the basic power production is traded on Elspot, but since changes will occur on a shorter timescale they can be traded on Elbas.

The grid-connected type of PV systems used in the study is installed on rooftops of private houses. The installation of these was a part of the project Sol300, which is now continued as Sol1000 [22]. Each PV system is connected via the electricity system of the household to the main electricity grid, and it has a peak power in the range of 1 to 4 kiloWatt-peak (kWp). The PV systems are spread all over Denmark (DK) and the total installed peak power of Sol300 is about 750 kWp [12] and of DK about 2.9 MWp. Other types of PV systems include PV power plants, where a large capacity of PV elements is installed in a small geographic location. Currently the Hoya de Los Vincentes, Jumilla, of Spain is the largest PV power plant with an installed peak power of 20 MWp. At the end of 2006 the world had a total of 5.8 GWp solar power and a growth rate of installed peak power in 2006 at 36% [26].

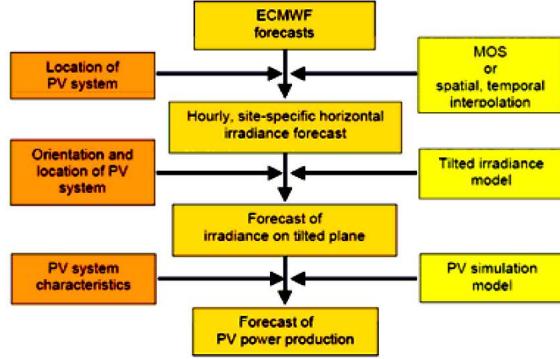


Figure 1.2: *Overview of forecasting method developed in recent German project [14]. The method is used for prediction of hourly values of power output from a PV system. NWPs are used as input and the method is used for horizons from 6 hours up to 3 days. The figure is taken from [14].*

1.2 Aim of the thesis

The primary aim of the thesis is to produce solar power forecasts to be used for trading at Elspot and little focus is on forecasts for trading at Elbas. This is done in order to narrow the scope of the thesis, but to give an indication of the performance at 1 to 11 hour horizons, they are included in the forecasts and evaluated on equal terms with the 12 to 36 hours horizons. A key point is to investigate whether prediction models using NWPs as input will perform better than prediction models that are solely based on past observations of solar power. Therefore two prediction model are developed: one using only past observations and one also utilizing NWPs. They are evaluated and compared to each other and to a persistence reference model.

1.3 Previous studies of prediction methods for solar power

Different areas in the field of PV studies have given useful information used in the present study. These are

- Parameterized models of power output from PV arrays. A comprehensive model has been developed at Sandia laboratories [13], where inputs such as direct irradiance, diffuse irradiance, air mass, temperature and wind speed, are used as inputs.

- Numerical weather prediction (NWP) models such as [4] are very extensive models of the atmosphere used to produce meteorological forecasts. A NWP model works on fields of atmospheric variables. To predict future states of the fields, the model is initialized using observed data and a set of mathematical equations for the physics and dynamics of the atmosphere are solved. The equations are impossible to solve exactly, so numerical methods are used. Studies investigating the optimization of NWP models with respect to global irradiance have been done by [6].
- Model output statistics (MOS) are statistical methods applied to the output of NWP models in order to predict other variables or to improve the predictions. MOS was first used to predict global irradiance in 1972 [7] and has recently been used by [9] to improve NWPs especially concerning local effects.
- Many approaches to forecasting of global irradiance have been developed. Two studies ([2] and [10]) from around 1990 use a transformation to remove the periodicity of the sun position and other deterministic information from time series of global irradiance and then linear models of the type ARIMA is used for prediction. Some later studies use solely different neural networks to predict the time series of global irradiance. Other very recent studies use satellite images or sky images from the ground for horizons less than 6 hours and NWPs for longer horizons. MOS is then performed by neural networks to improve the results.

The recent German project [14] is done with the same motivation as the present study. This is an extensive project including studies in all areas mentioned above. The project involves results both for single PV systems and ensembles of PV systems covering the entire Germany. The method developed uses satellite images for horizons less than 6 hours and NWPs for horizons up to 3 days. Figure 1.2 shows an overview of the method using NWPs.

Since it is found that the PV system power output is almost linearly dependent on global irradiance, the methods for prediction of global irradiance, can be directly applied to prediction of solar power. These methods are described more thoroughly in Chapter 2.

Finally it is noted that parts of the developed method are inspired by the Wind Power Prediction Tool (WPPT) developed at DTU [5].

1.4 Thesis outline

The development of the forecasting method is described progressively and the report should be read in sequence. The following is a short description of the chapters

Global irradiance prediction methods Methods for prediction of global irradiance developed in previous studies are described. This is done to give the reader an overview of how the prediction problem can be solved.

Data The data used in the present study is described and an exploratory analysis is carried out. The preprocessing of data before it is used in analysis is also described here.

Clear sky model The stochastic process of solar power output is not stationary. The distribution changes over the day due to the position of the sun. A developed method for transforming the process to a stationary process is described here.

Modeling procedure The details of the modeling procedure is described. Among other things it is defined how performance evaluation is done, the method used for fitting adaptive time series models is derived and a persistence reference model is described.

Autoregressive models A prediction model that only uses observations of the power production is developed. It contains an adaptive linear autoregressive model for each horizon.

Autoregressive models with exogenous input A prediction model that uses both observations of solar power and NWPs as input is developed.

Final evaluation and discussion The performance of the two developed models are evaluated and compared to the reference model. A discussion of the models and ideas to further work are outlined.

1.5 Notation

In this section some technical terms used in the text are explained. This should be looked through by the reader to clarify the meaning of these terms. Then a list of the mathematical symbols and a list of reserved symbols used in the report are given. For each reserved symbol there is a reference to its definition. Glossaries and an overview of symbols used in solar engineering are found in [20] and [24].

1.5.1 Clarification of terms used in the report

The following is a clarification of some terms used in the report.

The term *solar power* is used to denote PV power output and is given in Watt units. Mainly it is a mean value of the PV power output over a period of time, usually of length 15 minutes or 1 hour.

A *PV cell* or a solar cell is the basic PV building block and it is usually producing about 1 or 2 watts of power. To produce more power they are connected together and put between two layers of glass to form a *PV module*. Several PV modules put together forms a *PV array*. The complete setup with cables, inverter, PV modules, etc. is called a *PV system*.

The term *global irradiance* are in other studies denoted by: global radiation, global solar irradiance etc. It is a measure of the rate of total incoming solar energy on a horizontal plane at the Earth's surface.

The phrase *very-short term horizons* denote horizons less than 6 hours and the phrase *next day horizons* denote horizons from 12 to 24 hours.

In the text the term *toward dawn and dusk* is used. This means the hours close to dusk and dawn and thus it is where the sun elevation is low. The term *around noon* means the hours around noon.

All time values are UTC if nothing else is specified. When *day of year* or just *day* is used, then 2006-01-01 is *day of year* equal to 0 and thus 2006-12-31 is equal to 363.

1.5.2 Notation and reserved symbols

A few notes to clarify the notation:

- t without a subscript is used strictly as an index into time series, but t_x with a subscript denotes a time value, e.g. $t_{\text{tod}, \hat{p}, t}$ is the *time of day* of the t 'th sample in \hat{p} .
- Subscripts that are part of the name of the variable, are written with roman fonts, e.g. in the symbol $t_{\text{tod}, \hat{p}, t}$ the tod, \hat{p} subscript is a part of the name and t is an index.
- Sometimes indexing is done with square brackets as $\hat{\tau}_{\text{mat}}[i, k] = \hat{\tau}_{\text{mat}, i, k}$ and $\hat{\tau}_{\text{mat}}[:, k]$ denotes the entire k 'th column.
- A hat over a variable denotes that it is an estimate, e.g. \hat{p} is an estimate of p .

Generally used symbols throughout the report:

i, j : indexes,

t : index in univariate time series,

t_{tod} : Variable denoting *time of day*. Subscribing as $t_{\text{tod}, p, t}$ denotes the *time of day* of the t 'th sample in the time series p .

t_{day} : Variable denoting *day of year*. Subscribing as $t_{\text{day}, p, t}$ denotes the *day of year* for the t 'th sample in the time series p . Note that $t_{\text{day}} = 0$ is 2006-01-01.

The following list is an overview of reserved symbols used in the report. It is noted that this does not apply for Chapter 2:

- P_i : time series of accumulated solar power of the i 'th PV system,
defined in eqn. (3.1) on page 15
- p_i : time series of solar power of the i 'th PV system,
defined in eqn. (3.4) on page 19
- p : time series with the mean solar power of 21 PV systems,
defined in eqn. (3.5) on page 20
- $\{P_t\}$: the stochastic process of solar power, where p is a realization of
 $\{P_t\}$, derived in of Chapter (4) on page 27
- \hat{p}_i : time series with the i 'th forecast of solar power,
defined in eqn. (5.1) on page 46
- $\hat{\mathbf{P}}$: the matrix with all the solar power forecasts,
defined in eqn. (5.2) on page 46
- \hat{G}_i : the i 'th NWP forecast of accumulated global irradiance,
difined in eqn. (3.2) on page 15
- \hat{g}_i : the i 'th NWP forecast of global irradiance,
defined in eqn. (3.2) on page 15
- $\omega_{se}, \omega_\alpha$: The sun elevation angle and the sun azimuth angle,
shown in Figure 4.1 on page 28.
- α_{cut} : The limit defining which clear sky estimates are removed,
defined in eqn. 4.2 in page 36.
- $\{\tau_{S,t}\}$: the stochastic process resembling the transmittance of the at-
mosphere, derived in Chapter 4 on page 27.
- τ : the time series with the transformed p ,
defined in eqn. (4.1) on page 35.
- $\hat{\tau}_{nwp}$: The NWPs of τ ,
defined in eqn. (7.2) on page 73.
- $\hat{\tau}_i$: the time series with the i 'th forecast of τ ,
defined eqn. (5.3) on page 47.
- $\hat{\tau}_{mat}$: the matrix with all forecasts of τ ,
defined in eqn. (5.4) on page 47.

CHAPTER 2

Global irradiance prediction methods

As stated before the power output of a PV system (solar power) is almost linearly dependent on the global irradiance at the location. The uncertainties of predicted values of global irradiance are thus reflected directly in the predicted values of solar power, and this is by far the most influential effect on the quality of the solar power forecasts. Hence the problem of predicting the PV system power output is essentially the same as predicting the global irradiation. Methods to predict the global irradiance is therefore of highest relevance for the present thesis and several approaches have been developed over the last couple of decades. This chapter simply describes several of the methods to give the reader an overview of the problem. The methods are presented in historical order.

2.1 Parameterized atmosphere simulation and ARIMA(p, d, q) models

The article [2] from 1987 describes a method developed for sub-hourly global irradiance forecasting. Time series with both 3 and 10 minute values of observed global irradiance are used. The global irradiance is decomposed into a deterministic clear sky component and a stochastic cloud cover component. The

transmittance τ through the atmosphere is defined as

$$\tau = \frac{G}{G_{\text{ext}}},$$

where G is global irradiance and G_{ext} is extraterrestrial irradiance which can be calculated. Then τ is decomposed into two components

$$\tau = \tau_a \cdot \tau_c$$

where τ_a is the transmittance of the atmosphere in clear sky and τ_c is the cloud cover component. τ_a is modeled by a parameterized atmosphere simulation model. The input parameters are among others: optical air mass, ozone path length, scattering albedo and sun-earth geometrical relationship. τ_c is Gaussian and it is modeled by ARIMA(p, d, q) models. The data is divided into periods of one week and the ARIMA models are identified using the ACF and the partial ACF, for each of the periods. This results in different models for each period. The τ_c time series is differenced one time, i.e. $d = 1$, the MA-part is mostly $q = 2$ and many different AR lags are included, e.g. one week period has AR lags 5,6,8,9,11 included and the next week has only AR lag 5 included. This is an overfitting of the data, but it indicates that the τ_c is hard to predict.

2.2 Identification of ARMA(p, q) model by the use of a Kalman filter

An autoregressive moving average (ARMA) model identification on transformed hourly time series of global irradiance is described in the article [10] from 1990. The global irradiance is transformed into a stationary process by a clear sky model. One clear sky model is used for the winter period and one for the summer period. The ARMA models are then identified by the use of a Kalman filter, a method proposed by H. Madsen in 1985, and it is left to the reader to look for further details of this method. The result found is an ARMA(3,1) for the summer period and an ARMA(2,2) for the winter period. No evaluation of the models is carried out.

2.3 Forecasting using artificial intelligence techniques

The article [21] from 2000 explores artificial intelligence (AI) techniques for prediction of global irradiance (G). Hour values of global irradiance is used and no

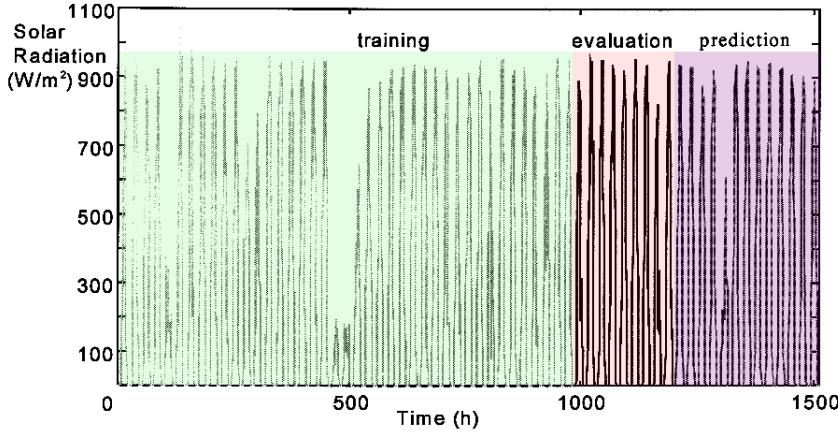


Figure 2.1: The global radiation data used in [21]. The training period is 41 days, the evaluation period is 13 days and the prediction period is 9 days.

transformation is necessary. Furthermore observations of temperature (T), pressure (P), wind speed (WS) and wind direction (WD) is used. Models with and without the additional climate variables are found and evaluated. The artificial intelligence techniques used are artificial neural networks and adaptive neuro-fuzzy inference systems. The conventional model is an ARIMA(1, 0, 1)(0, 1, 1)₂₄ model and a linear regression (LR) model

$$\begin{aligned} G(t|t-1) = & a_1G(t-1) + a_2G(t-2) + a_3G(t-3) + b_1G(t-24) + c_1T(t-1) \\ & + c_2T(t-6) + d_1WD(t-1) + e(t). \end{aligned}$$

It is noted that the ARIMA model is applied with only a seasonal differencing and the resulting process is obviously not stationary. The result of one-step prediction is evaluated by root mean square error (RMSE) improvement over a persistence model. For the ARIMA model it is found to be 62.89% RMSE improvement and the best performing AI technique without additional inputs has a RMSE improvement of 71.94 %. Including the additional variables gives the LR model 60.91 % and the best AI model 74.04 %.

The results from this study must be criticized for at least two things. Applying the ARIMA and LR models on a non stationary process will decrease their performance and thus the comparison of the AI models and the conventional models are problematic. Secondly the size of the dataset used is very sparse. The solar radiation data used in the study is shown in Figure 2.1. Considering that 8 out of the 9 days in the prediction period are clear sky days, without any noticeable randomness, this dataset seems to be very sparse. Though it is commented in the conclusion that the dataset exhibit a typical number of clear

sky days for the region, the use of only 9 days is a small dataset to conclude upon.

2.4 Prediction using sky images

A recent study [3] from 2006 use artificial neural networks and cloudiness indices. The study utilizes a Total Sky Imager which makes sky images every minute. From the sky image three clearness indices (a ratio from 0 to 1) are calculated for each pixel: clear sky, high density clouds and low density clouds.

The artificial neural networks uses this together with observations of global irradiance as input. Iterative predictions of up to 1 hour (60 steps) are made. The results of the study is that the estimation of the cloudiness indices still have too high errors for the method to work satisfactorily.

2.5 Using satellite images and NWPs

The study that in terms of type of forecasts produced and aim is most similar to the present study, is described in [9] from 2006 and [14] from 2007. These articles outline the results of several studies done at Oldenburg University, Germany. The relevance is increased by the fact that Germany has an installed PV effect of 2.9 GW (in the end of 2006, [26]), which is more than half of the global installed PV effect. Furthermore power produced in the northern part of Germany is also traded on Elspot and Elbas.

Two methods are developed for global irradiance forecasting: A very-short term method for horizons below 6 hours where satellite images is used, and a short-term method for horizons up to 3 days where NWPs are used. The following first describes the very short-term method.

Satellite images are observations in 2D of solar radiation reflected in the atmosphere. A cloud cover index is calculated for each position (pixel) in the image using the Heliosat method [1]. This is a ratio measuring the cloud cover at the position of the pixel. In the very short-term method, images from the geostationary METEOSAT satellite is used, and these are retrieved every 15 minutes and have a resolution of approximately 1 km x 1 km. The images are forecasted by calculating a motion vector field from a sequence of images, and this is used to project the most recent image into a forecast image of a given horizon. The forecasted image is finally smoothed and the global irradiance for each position in the image is calculated. Figure 2.2 show an overview of the method.

Another method has been developed for forecasting hour values up to 3 days

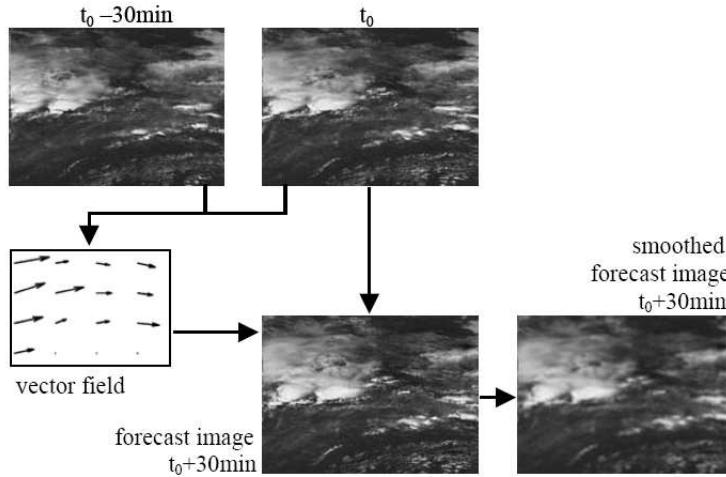


Figure 2.2: *Overview of the very short-term method developed in Oldenburg University. Satellite images are forecasted by calculating a motion vector field. The method is used for horizons less than 6 hours. The figure is taken from [9].*

ahead. This method is based on NWPs. The performance of several NWP models of different spatial resolution have been investigated. The global model run by European Centre for Medium-Range Weather Forecasts (ECMWF), which has a temporal resolution of 3 hours and a spatial resolution of 25 km x 25 km, used with temporal and spatial interpolation showed the best performance in the evaluations. MM5 is a mesoscale NWP model that is initiated with output from global models. A study has been carried out, where seven configurations of MM5 with a spatial resolution of 3 km x 3 km was evaluated for performance regarding global irradiance prediction. It was found that there is a high deviation between the performance of the different configurations. Finally the methods where model output statistics (MOS) where applied showed promising results. In the used MOS, different NWP output is used as input to a statistical model, which predicts solar irradiance. The NWP outputs of most relevance was found to be: cloud cover index, dew point difference, 500 hPa relative humidity, cloud cover below 2000 m and probability of precipitation. Artificial neural networks were used for the statistical models.

The studies are also concerned with forecast accuracy. The satellite method classifies the cloud index images into different weather situations. Forecasts in clear sky have a higher accuracy, than forecasts with broken clouds. The solar zenith angle is also taken into account.

Finally the solar power is modeled using the forecasted global irradiance. Figure 1.2 showed in the introduction chapter, gives an overview of this model. First the

(horizontal) global irradiance is transformed into irradiance in the tilted plane of the PV system, and then a PV simulation model is applied to get the solar power. As mentioned the developed methods were evaluated on ensembles of PV systems scattered all over Germany. The results showed that the reduction factor

$$r = \frac{RMSE_{\text{ensemble}}}{RMSE_{\text{single}}}$$

depends of the size of the region where the PV systems are uniformly distributed within. For the entire Germany a reduction factor of about 0.4-0.5 is reported.

CHAPTER 3

Data

The data used in the study comes from three sources:

Sol300 Observations of power from PV systems (solar power) installed at eight locations in Denmark. The data cover the period from 2006-01-01 to 2007-01-01. It is from the project Sol300 [22] which is run by the Danish company Energimidt. The original dataset consists time series with 15 minute values of accumulated power production a PV system. There are 182 time series. The time series for the i 'th PV system is

$$P_i = \{P_{i,t}; t = 1, \dots, 35040\}. \quad (3.1)$$

Danish Metheological Institute Forecasts of global irradiance and eight other variables. Danish Metheological Institute (DMI) uses the numerical weather prediction (NWP) model DMI-Hirlam [4]. The forecasts have a 48-hour horizon and are updated every 12 hour at 00:00 UTC and 12:00 UTC, the first forecast is at 2006-01-01 00:00 and the last is from 2007-31-12 12:00. Each forecast is a time series with 3 hour values of accumulated global irradiance. The i 'th forecast is

$$\hat{G}_i = \{\hat{G}_{i,k}; k = 1, \dots, 16\} \quad (3.2)$$

Power production and irradiance observations Data with observations of power production from solar cells, irradiance perpendicular to the solar cell plane and horizontal diffuse irradiance. 15 minute values have been

measured at 16 locations well spread over DK. The used data are from the period 2007-03-22 to 2007-10-30. This data is only used to show the relation of these three quantities and are not used in the prediction models.

The following sections describe data and operations applied to get a synchronized and well scaled dataset. An exploratory data analysis is made to see some general trends, for example a higher power production in the summer period than in the winter period should be observed.

3.1 Scaling and synchronization of data

All the data is scaled and transformed into units that are equivalent: solar power into Watts [W] and irradiance into Watts per square meter [$\frac{W}{m^2}$].

This operation implies differencing, and to keep the correct synchronization when differencing a time series, e.g. the power of the i 'th PV system

$$P_i = \{P_{i,t}; t = 1, \dots, N\}$$

into p_i . Then the time value $t_{P_i,t}$ of the t 'th sample in P_i , must be corrected by

$$t_{p_i,t} = t_{P_i,t} - \frac{1}{2} t_{sp}$$

where t_{sp} is the sample period and $t_{p_i,t}$ is the time value of the t 'th sample in the differenced time series. It is noted that the horizon for predictions at t of the differenced time series still is $k \cdot t_{sp}$, but the time value of the k 'th horizon has been modified.

Furthermore it is noted that when resampling time series into another frequency the time values have to be modified to keep correct synchronization. For example when resampling a time series y_q with 15 minutes values into a time series y_h with hour values by

$$y_{h,i} = \frac{1}{4} \sum_{j=4(i-1)+1}^{4i} y_{q,j}$$

the time value are modified by

$$t_{y_h,i} = t_{y_q,4i} - \frac{3}{2} \cdot 15 \text{ min.}$$

In order to use all the data in the same analysis the time reference for all the data is shifted to Coordinated Universal Time (UTC). The time reference of the Sol300 data is Central European Time (CET) and two steps are applied to shift it into UTC:

CEST to CET In the summer period the clock is moved one hour ahead and Central European Summer Time (CEST) is used for the Sol300 data. To synchronize the data in this period one hour is subtracted from time values between 2006-03-26 01:00 UTC and 2006-10-29 01:00 UTC.

CET to UTC One hour is subtracted from the time values.

In the report all time and date values are in UTC if nothing else is specified.

After performing these operations, synchronization errors still existed in the Sol300 data. To reveal these errors the P_i before any of the synchronization steps, is differenced into

$$p_{i,t} = P_{i,t} - P_{i,t-1}. \quad (3.3)$$

Then the time value of all positive values of $p_{i,t}$ for all i and t are plotted. The time values as a function of *day* and *tod* (the *time of day*) are shown in Figure 3.1.

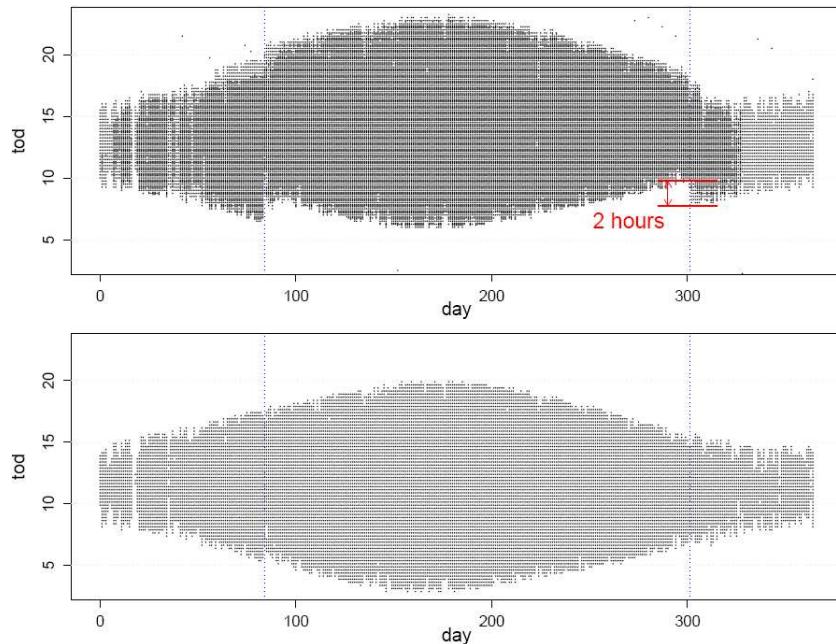


Figure 3.1: Time values of samples with positive values of p . Sudden jumps in the border indicates a shifted time reference or synchronization errors. Upper plot: Samples where $p_{i,t} > 0$ for all i and t . $p_{i,t}$ is defined in (3.3). Lower plot: The synchronized and error checked time series of solar power p (defined in eqn. (3.5)). No obvious synchronization errors in p are observed.

If the border of these dots have sudden jumps, this indicates synchronization shifts. The upper plot shows the original data before any synchronization, only a differencing is applied. The 2 hour gap marked, show that the data apparently is shifted +2 hours. This is an error in the dataset, since the shift should only be +1 hour in the summer period. This must be caused by errors in the system that originally made the data. Otherwise it could have happened in the process of exporting the data from the given database of observations. Apart from this synchronization error, it is observed that a few outliers exists. The lower plot shows the synchronized and error checked time series p defined in eqn. (3.5). When comparing the two plots, it is seen that the density of the dots are higher in the upper plot. This indicates that many of the $p_{i,t}$ are not synchronized with the others. Finally it is seen that when the Sol300 data has been synchronized and error corrected, no obvious synchronization errors can be observed in the plot.

3.2 Sol300

Each P_i in the Sol300 dataset is from a PV system with installed effect in the range of 1 to 4 kiloWatt-peak. The PV systems are located at 8 different locations in DK. Since the scope of the thesis doesn't include spatial effects, only data from the location Brædstrup - that has 52 PV systems, which is the highest number - is used. All P_i from Brædstrup are shown in Figure 3.2 and it can be seen that data from some of the PV systems does not cover the whole year. That is, they contain missing values, and hence some error correction is needed.

In order to make complete time series with no missing values, the following steps are applied to the data:

1. P_i that have periods with more than 1000 succeeding samples (approximately 10 days) that are missing values, are removed.
2. Periods of missing values from P_{i,t_s} to P_{i,t_e} and where $P_{i,t_s-1} = P_{i,t_e+1}$ (i.e. no production has occurred in the period) are filled in by setting $P_{i,t} = P_{i,t_s-1}$ for $t = t_s, \dots, t_e$.
3. All P_i that still have missing values are removed.
4. P_i with periods of no power production for longer than 4 days are removed.
This is excepted for January and December, since many days in these two months have no solar power production.

After these steps 25 P_i remained in the set, but further error correction was necessary since errors still existed in the data.

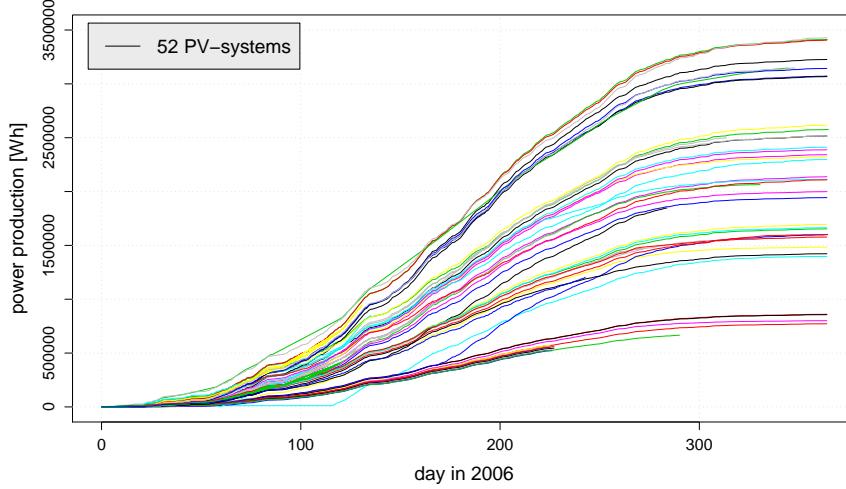


Figure 3.2: P_i from each of the 52 PV systems in Brædstrup. In the plot it is seen that the PV systems have different installed effect and that there are missing values.

Then scaling was applied. The P_i are given in [Wh] with a sample period of 15 minutes and are transformed into [W]. The transformed power unit is

$$\left[\frac{\text{Wh}}{\frac{1}{4}\text{h}} \right] = [4\text{W}] = 4 \cdot [\text{W}]$$

The transformation from [Wh] to [W] is

$$p_{i,t} = 4 \cdot (P_{i,t} - P_{i,t-1})$$

and it is noted that $p_{i,1} = P_{i,1}$. So the time series

$$p_i = \{p_{i,t}; t = 1, \dots, 35040\} \quad (3.4)$$

is the differenced accumulated power production of the i 'th PV system.

Closer investigations of p_i revealed synchronization errors between the PV systems in the dataset. Figure 3.3 shows

$$\dot{p}_{t_{\text{day}},i} = \frac{p_{t_{\text{day}},i}}{\max_i(p_{t_{\text{day}},i})}$$

where $p_{t_{\text{day}},i}$ is p_i for the day of the year t_{day} and $\max_i(p_{t_{\text{day}},i})$ is the maximum value of $p_{t_{\text{day}},i}$ with respect to i . Note that $t_{\text{day}} = 0$ is the 1st of January. The plot show all 25 $\dot{p}_{t_{\text{day}},i}$ for the second day after the shift to summer time, $t_{\text{day}} = 85$. It is seen that for $i = 15, \dots, 26$ $\dot{p}_{85,i}$ is one hour delayed. This

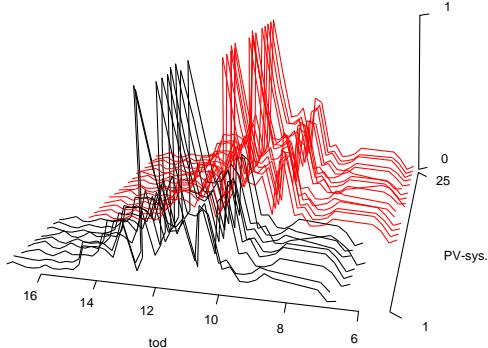


Figure 3.3: Plot of $\dot{p}_{85,i}$ for all PV systems. Day 85 is the second day of the summer time period, 2006-03-27. The data is from the same day since it is seen that the pattern of all lines are almost identical, but the red lines are clearly one hour delayed compared to the black lines.

occurred on 3 days hereafter where a decreasing number of $\dot{p}_{t_{\text{day}},i}$ were delayed. The similar errors occurred in exactly the same manner 5 days after the shift to normal time. This led to a manual check of all days and also other errors were found, resulting in the removal of 4 more p_i from the dataset. Following the manual check, a computational check of the time values for equidistance was done. Finally all separate days with errors were manually checked again concluding that no obvious errors exist in the data.

The final dataset used in the analysis consists of 21 p_i from PV systems at the location Brædstrup, which have been thoroughly checked for errors. In the rest of the report the time series

$$p = \{p_t; t = 1, \dots, N\} \quad (3.5)$$

where

$$p_t = \frac{1}{21} \sum_{i=1}^{21} p_{i,t}$$

is used for analysis. p is used in three sample rates:

- With sample period $t_{\text{sp}} = 15$ minutes, number of samples $N = 35040$ and the time value of the first sample is $t_1 = 2006-1-1 00:07:30$.
- $t_{\text{sp}} = 1$ hour, $N = 8760$ and $t_1 = 2006-01-01 00:30$.
- $t_{\text{sp}} = 3$ hours, $N = 2920$ and $t_1 = 2006-01-01 01:30$.

The time values of the samples in p are

$$t_{p,t} = t_1 + t \cdot t_{\text{sp}}.$$

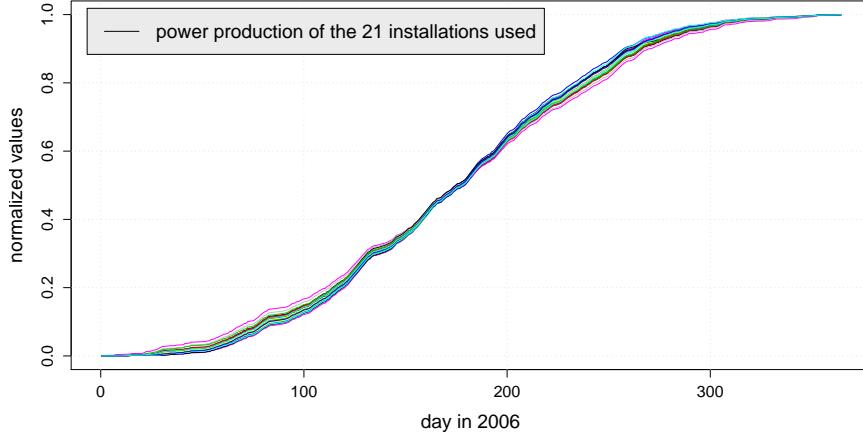


Figure 3.4: The final dataset used in the remaining of the analysis, consists of time series from 21 PV systems in location Brædstrup. The plot shows \dot{p}_i , which as expected show that the highest solar power production is in the summer period and have the same local trends.

If nothing else is specified then p with 15 minute values is used in this chapter and chapter 4 Clear Sky model, and p in hour values is used in all other parts of the report.

The final dataset is shown in Figure 3.4 by a plot of

$$\dot{p}_i = \left\{ \frac{\sum_{j=1}^t p_{i,j}}{\sum_{j=1}^{35040} p_{i,j}}, t = 1, \dots, 35040 \right\}.$$

It is seen that all \dot{p}_i have the same trends and that some have a relatively higher production in the spring than in the autumn - which is attributed to changes in the PV systems.

3.3 DMI

Forecasts of climate variables are provided by DMI. As mentioned the NWP model Hirlam [4] has been used. They are updated daily at 00:00 and 12:00 UTC and each forecast contains the following horizons +3h, +6h, ..., +48h. The observations of the variables at the time of update are also given. Predictions of the following variables are given: Global irradiance (given as accumulated values and the i 'th forecast of these is a time series denoted by \hat{G}_i), Fog, Total cloud cover, Low cloud cover, Medium cloud cover, High cloud cover, Temperature

(the temperature 2 m. above the ground), Wind speed and Wind direction. The fog and cloud cover variables are given in values 0, 0.1, ..., 1.

All these variables have a potential for predicting solar power [13]. Due to limiting the scope of the report, only \hat{G}_i is used.

The \hat{G}_i is given in $[\frac{J}{m^2}]$ with sample period of 3 hours and it is transformed into $[\frac{W}{m^2}]$ units. The transformed unit is equal to

$$\left[\frac{\frac{J}{m^2}}{3h} \right] = \left[\frac{1}{3600} \cdot \frac{\frac{Wh}{m^2}}{3h} \right] = \frac{1}{3 \cdot 3600} \cdot \left[\frac{W}{m^2} \right]$$

so the transformation from $[\frac{J}{m^2}]$ to $[\frac{W}{m^2}]$ is

$$\hat{g}_{i,k} = \frac{1}{3 \cdot 3600} \cdot (\hat{G}_{i,k} - \hat{G}_{i,k-1}).$$

So the time series

$$\hat{g}_i = \{\hat{g}_{i,k}; k = 1, \dots, 16\}.$$

is the i 'th forecast of differenced global irradiance. The time value where the i 'th prediction is given is

$$t_{\hat{g},i} = t_{\hat{g},1} + (i - 1) \cdot 12h$$

where $t_{\hat{g},1} = 2006-01-01 00:00$. The time value of $g_{i,k}$ is

$$t_{\hat{g},i,k} = t_{\hat{g},i} + k \cdot 3h - 1.5h$$

where 1.5h are subtracted because of the differencing.

Investigation of the forecast data reveals errors in 6 forecasts where values are missing for one or two succeeding time values. These missing values have been replaced by linearly inter- and extrapolated values using the two nearest samples. It is noted that it was checked manually that the extrapolated values are within the minimum and maximum limits of the other values in the forecast.

To get an overview data an exploratory analysis is done by making plots showing \hat{g}_i compared to the solar power. Figure 3.5 shows four \hat{g}_i transformed by the linear regression line showed in Figure 3.7, and the main trend in \hat{g}_i seems to follow p well.

Time series with 3 hour values for a given j are made by

$$\hat{g}_{k,j} = \{g_{1,j}, \dots, g_{1,j+4}, g_{2,j}, \dots, g_{2,j+4}, \dots, g_{n,j}, \dots, g_{n,j+4}\}.$$

For $j = 1$ this is the time series of the most recent updated predictions, containing the values for horizons $k = 1, \dots, 4$, for $j = 5$ this is the most recent

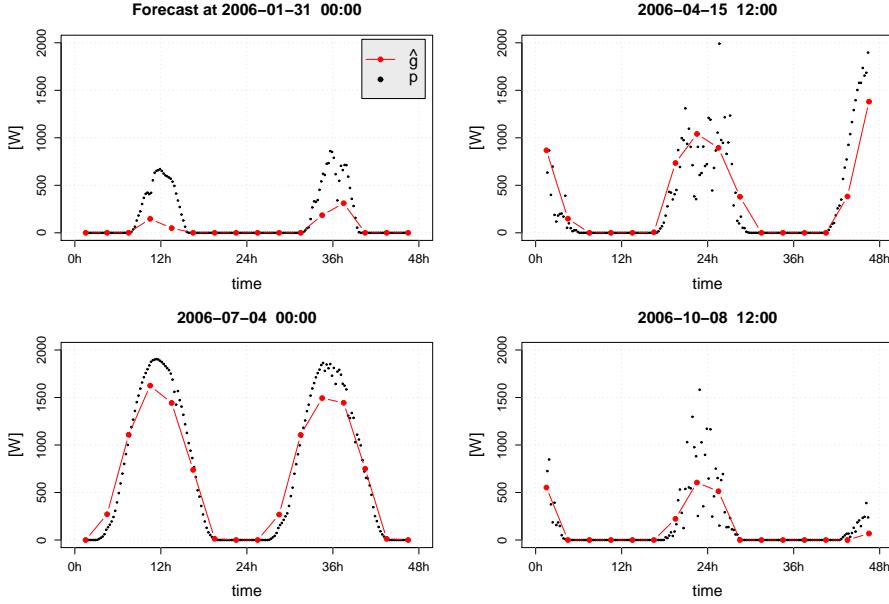


Figure 3.5: Plots of four \hat{g}_i and corresponding values of p from different periods of the year. The trends of \hat{g}_i agree with the main trends in p , but can obviously not describe fluctuations at frequencies higher than $\frac{1}{2} f_{\hat{g}} = \frac{1}{6h}$.

updates of horizons $k = 5, \dots, 8$ etc. To see if the annual trend of p is followed by $\hat{g}_{k,j}$, the year is divided into two week periods, i.e. 26 periods, where the mean of each period is made for both p and $\hat{g}_{k,j}$. They are now normalized by dividing with their maximum value and plotted in Figure 3.6. Similar plots are checked for all horizons, i.e. $j = 1, 5, 9, 13$ and all plots show that the annual trend in p is followed well by $\hat{g}_{k,j}$.

The time series with 24 hour values for a given horizon k

$$\hat{g}_{00,k} = \left\{ \hat{g}_{2t-1,k}; t = 1, \dots, \frac{n}{2} \right\}$$

is the predictions made at 00:00 of horizon k (since \hat{g}_i is in 3 hour values the time of the k 'th horizon is here $k \cdot 3$ hours). The similar time series for predictions made at 12:00 is

$$\hat{g}_{12,k} = \left\{ \hat{g}_{2t,k}; t = 1, \dots, \frac{n}{2} \right\}.$$

Using p resampled to 3 hour values the time series with 24 hour values

$$p_{13:30} = \{p_{(24(t-1)+5)}; t = 1, \dots, 364\}$$

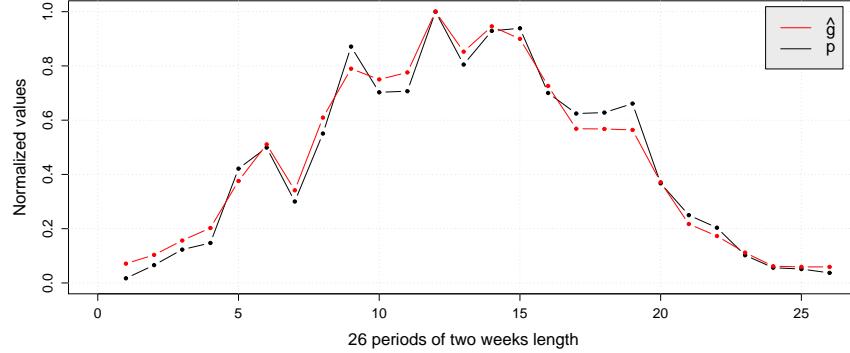


Figure 3.6: The mean of two week periods for $\hat{g}_{k,1}$ and p , both normalized by dividing with their maximum value. They have the same annual trend and fluctuations in p are followed well by $\hat{g}_{k,1}$.

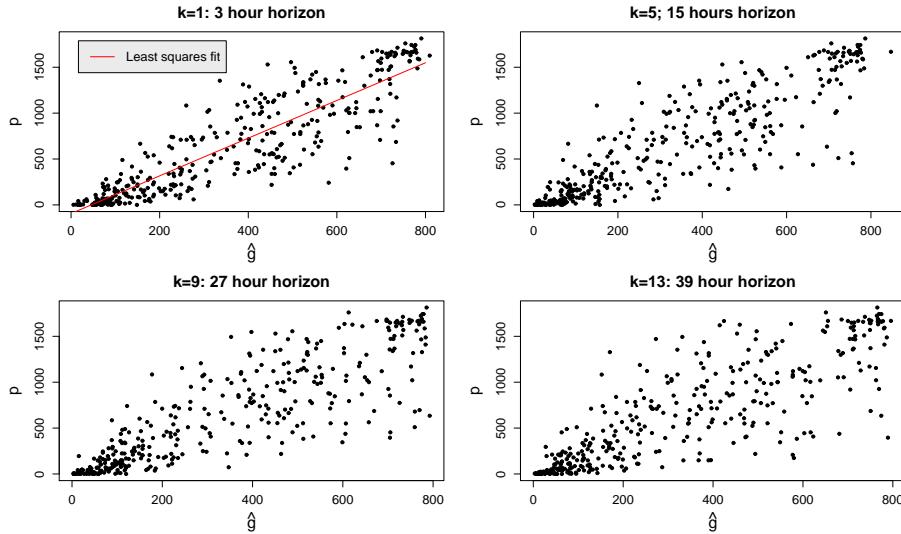


Figure 3.7: Plots of $p_{13:30}$ as a function of: The upper-left $\hat{g}_{12,1}$, the upper-right $\hat{g}_{00,5}$, the lower-left $\hat{g}_{12,9}$ and the lower-right $\hat{g}_{00,13}$. The plots indicate that there is a dependency between the NWPs of global irradiance and the power production of the PV systems. It is seen that for high and low values of $\hat{g}_{t_i,k}$ the errors are smaller than for $\hat{g}_{t_i,k}$ in the middle of its range, and that the errors seems to be increasing for increasing horizon.

is the solar power at 13:30 each day. To see if there is a dependency between the solar power and the predictions of global irradiance, plots of $p_{13:30}$ as a function of respectively $\hat{g}_{12,1}$, $\hat{g}_{00,5}$, $\hat{g}_{12,9}$ and $\hat{g}_{00,13}$ is shown in Figure 3.7. It is found that there is clear indication of dependency, and this was of course expected. Similar plots of all other $\hat{g}_{00,k}$ and $\hat{g}_{12,k}$ with their corresponding $p_{t_{\text{tod}}}$ have been checked to find outliers and other errors.

3.4 Power production and irradiance observations

The Danish company Energimidt is in cooperation with Energinet.dk measuring the three quantities: power production of a PV system (a PV system with only one PV module), global irradiance and solar irradiance incident on the PV module surface I_{surface} . This is done at 16 locations in DK with 15 minute values. It must be noted that this data was available through a website, where only values for one day at a time could be downloaded wrapped in html. Data was acquired for location Brædstrup between 2007-03-22 and 2007-10-30. The irradiance observations are done with silicon irradiance sensors of the type Mencke & Tegtmeier Si-01TCext-K. This is a small standard silicon PV element with temperature compensation. One is placed horizontally measuring the global irradiance.

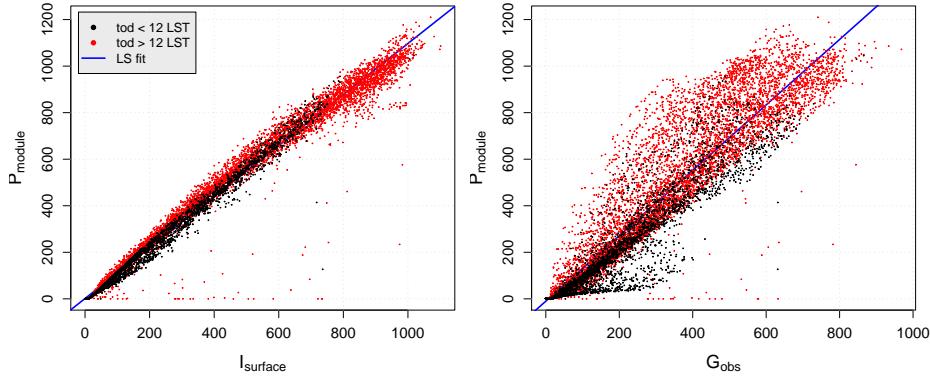


Figure 3.8: Left: solar power production versus solar irradiance incident on the PV module surface I_{surface} . An clear indication of an almost linear dependency is seen. Right: solar power production versus global irradiance. An indication of a linear dependency is seen, though with a higher variance around the LS fit than for solar power production versus I_{surface} . The time reference is local solar time (LST).

Another is placed pointing the same direction as the PV module, and is measuring the irradiation incident to the surface. Figure 3.8 show plots of the solar power versus global irradiance and versus I_{surface} . An indication of a linear dependency is seen. The variation of the distribution around the least squares (LS) fit is smaller for the solar power versus I_{surface} , and this seems very reasonable. In the left plot there seems to be a tendency that the ratio $\frac{\text{solar power}}{\text{global irradiance}}$ is higher for $t_{\text{tod}} > 12$, but this effect is not seen for $\frac{\text{solar power}}{I_{\text{surface}}}$. This indicate that a higher share of the global irradiance strikes the surface in the afternoon, and thus it is found that the PV module is pointing westward. Furthermore this effect should be accounted when the function between global irradiance and solar power is modeled, this is done in Section 7.1. It is noted that the data described in this section is only used in the modelling of the function between global irradiance and solar power.

CHAPTER 4

Clear sky model

The solar power time series $p = \{p_t, t = 1, \dots, N\}$ with 15 minute values and $N = 35040$, is the realization of a stochastic process $\{P_t, t = 1, \dots, N\}$. In order to model $\{P_t\}$ effectively with conventional time series models, the characteristics of the process must be examined. The most conventional models and methods assume stationarity [15]. This is not fulfilled for $\{P_t\}$ since its first moment is dependent of *time of day*, $t_{\text{tod},t}$. The n -simensional probability distribution function is thus not invariant to changes in time and the process is not stationary. Hence a transformation of $\{P_t\}$ to obtain a stationary process is needed.

The transmittance of the atmosphere τ_{atm} is defined as

$$\tau_{\text{atm}} = \frac{G}{G_{\text{ext}}}$$

where G_{ext} is extraterrestrial and G is global irradiance, i.e. the irradiation on a horizontal surface at the surface of the earth. A similar quantity $n_{i,j}$ is called the cloud cover index [1] and is a ratio calculated using the albedo measured in the (i, j) pixel of a satellite image. It is an estimate of the cloud cover percentage at the pixel. $n_{i,j} = 0$ indicates no clouds (clear sky) at the position of the pixel and $n_{i,j} = 1$ indicates maximum cloud cover.

$\{P_t\}$ is divided into a deterministic and a stochastic part

$$P_t = p_{\text{cs}}(t_{\text{p},t}) \cdot \tau_{\text{S},t}$$

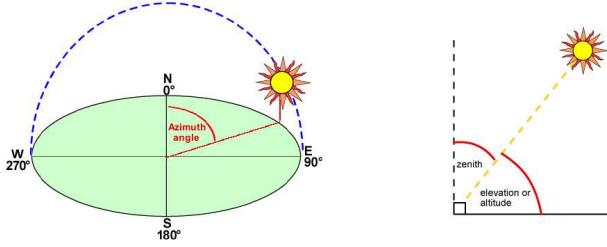


Figure 4.1: The sun elevation angle ω_{se} and the sun azimuth angle ω_α . The figure is taken from [11].

where $t_{p,t}$ are the time value of the t 'th sample in p , $p_{cs}(t_{time})$ is a deterministic function which is defined so that when there is no cloud cover, i.e. clear sky, then $p_{cs}(t_{time}) = P_t$. The $\{\tau_{S,t}\}$ is a stochastic process. It is not exactly one of the two above mentioned transmittance and cloud cover quantities, but very alike. It is closer to satisfying the assumption of stationarity than $\{P_t\}$, this is showed in Section 4.5.

The position of the sun is a deterministic function of time and it is periodic over one year. Therefore $p_{cs}(t_{time})$ can be modeled by either

$$\begin{aligned} p_{cs}(\omega_{se}, \omega_\alpha), \\ p_{cs}(t_{day}, t_{tod}) \end{aligned}$$

where ω_{se} is the sun elevation and ω_α is the sun azimuth, shown in Figure 4.1, and t_{day} is the *day of year* and t_{tod} is the *time of day*. Note that these four variables are really functions of time t_{time} , but this have been left out of the notation and is assumed implicit. Definitions of ω_{se} and ω_α and equations for calculating them as a function of *time of year* are found in [11].

The transformation of $\{P_t\}$ into $\{\tau_{S,t}\}$ is done by

$$\tau_{S,t} = \frac{P_t}{p_{cs}(t_{p,t})} .$$

Hence to make the transformation a method for estimation of $p_{cs}(t_{time})$ is needed. In clear sky where

$$\tau_{S,t} = 1 \Rightarrow p_{cs}(t_{p,t}) = P_t$$

then $p_{cs}(t_{time})$ can be observed directly from $\{P_t\}$, but (unfortunately) this sky condition does not occur all the time. The method for estimation of $p_{cs}(t_{time})$ needs to work when $\{\tau_{S,t}\}$ is varying.

In exploring p to find out how $p_{cs}(t_{time})$ can be modeled, a couple of plots are made to see if the function changes over time and to find out if ω_{se} is enough

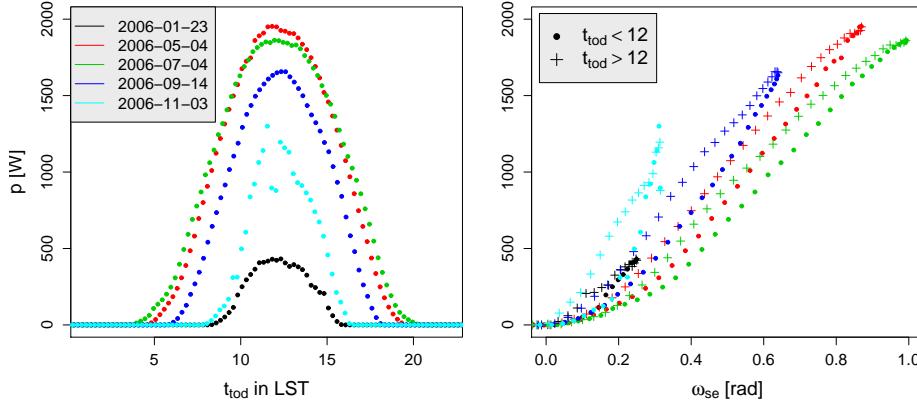


Figure 4.2: 5 clear sky days spread over the year. t_{tod} is in LST. Left: p versus t_{tod} . Right: p versus ω_{se} . A clear change over the year and also different levels between p_{am} and p_{pm} is observed.

input to the model. Local Solar Time (LST) is defined so that at 12:00 LST then $\omega_\alpha = \pi$ i.e. the sun is due south at the given longitude. At Brædstrup

$$t_{\text{UTC}} \approx t_{\text{LST}} - 38 \text{ min}$$

so at $t_{\text{tod}} \approx 11:20$, the sun is due south. This leads to Figure 4.2, which show 5 clear sky days distributed over the year. It is seen in the left plot that the curve and thus also $p_{\text{cs}}(t_{\text{time}})$ is changing over the year. The right plot shows that the curve is also different for $p_{\text{am}} = p[t_{\text{tod}} < 11:20]$ - that is the values in p before LST noon - than for $p_{\text{pm}} = p[t_{\text{tod}} > 11:20]$. It is found that $p_{\text{cs}}(t_{\text{time}})$ changes over the year and that ω_{se} alone is not sufficient to find the best model of $p_{\text{cs}}(t_{\text{time}})$, since it is also dependent of ω_α .

To explore this further two statistics

$$\bar{p}_{\text{am},i} = \frac{1}{n_{\text{am}}} \sum_{i=1} n_{\text{am}} p_i[t_{\text{tod}} < 11:20]$$

and

$$\bar{p}_{\text{pm},i} = \frac{1}{n_{\text{pm}}} \sum_{i=1} n_{\text{pm}} p_i[t_{\text{tod}} > 11:20],$$

are calculated for each p_i . $\bar{p}_{\text{am},i}$ is the mean of the solar power of the i 'th PV system before LST noon, and $\bar{p}_{\text{pm},i}$ is after LST noon. The model

$$\frac{\bar{p}_{i\text{pm}}}{\bar{p}_{i\text{am}}} = \alpha a_{PV_i} + \epsilon_i, \quad \epsilon \sim N(0, \sigma^2)$$

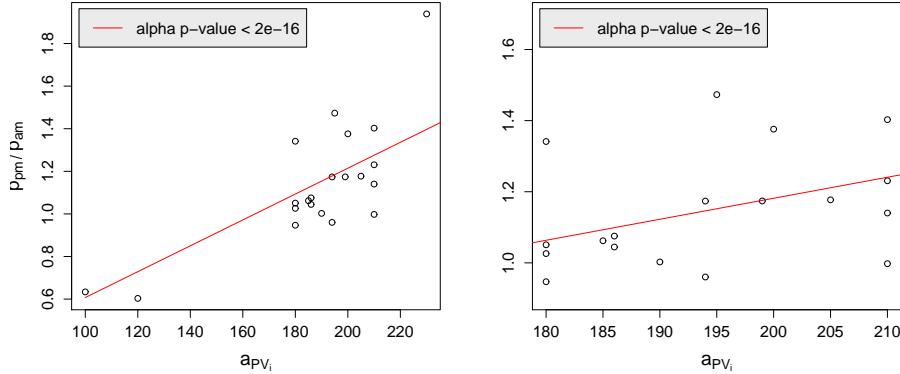


Figure 4.3: $\frac{\bar{p}_{i,\text{pm}}}{\bar{p}_{i,\text{am}}}$ as a function of a_{PV_i} the azimuth of the i 'th PV systems. The line shows a LS linear regression. Left: All 21 PV systems. Right: Without the 3 outliers. Both fits have significant p -value for α .

is fitted by LS and it is found that for a given PV system there is a significant correlation between $\frac{\bar{p}_{\text{pm}}}{\bar{p}_{\text{am}}}$ and the azimuth of the PV system. Figure 4.3 show the results. It leads to conclude that a PV system pointing westward will produce relatively more electricity in the afternoon, than a PV system pointing eastward at the geographic location of Bræstrup. It is therefore found that ω_{se} is not sufficient input for a model of $p_{\text{cs}}(t_{\text{time}})$, since $p_{\text{cs}}(t_{\text{time}})$ is also dependent on ω_{α} .

It must be noted that since $P_t = 0$ when the earth is blocking the solar irradiation, then $\{\tau_{S,t}\}$ can not be observed at nighttime. Furthermore note that toward dawn and dusk where $p_{\text{cs}}(t_{\text{time}}) \rightarrow 0$, then noise in p and errors in the estimates of $p_{\text{cs}}(t_{\text{time}})$ will have an increasing effect on estimates of $\{\tau_{S,t}\}$.

Some technique for interpolating the curve of $\{P_t\}$ between clear sky days is needed to estimate $p_{\text{cs}}(t_{\text{time}})$. Figure 4.4 shows p versus t_{day} and t_{tod} . This space - spanned by t_{day} , t_{tod} and $\{P_t\}$ - it is referred to as the *time space*. The values form a bell shaped point cloud. Figure 4.5 shows p as a function of ω_{se} and ω_{α} . This space - spanned by ω_{se} , ω_{α} and $\{P_t\}$ - is referred to as the *sun position space*. The upper surface of the point clouds, which intercepts p at clear sky days, is an estimate of $p_{\text{cs}}(t_{\text{time}})$. The remainder of this chapter first describes two methods for estimating the surface and then three results of estimating $p_{\text{cs}}(t_{\text{time}})$, of which two are made in the *time space* and one in the *sun position space*.

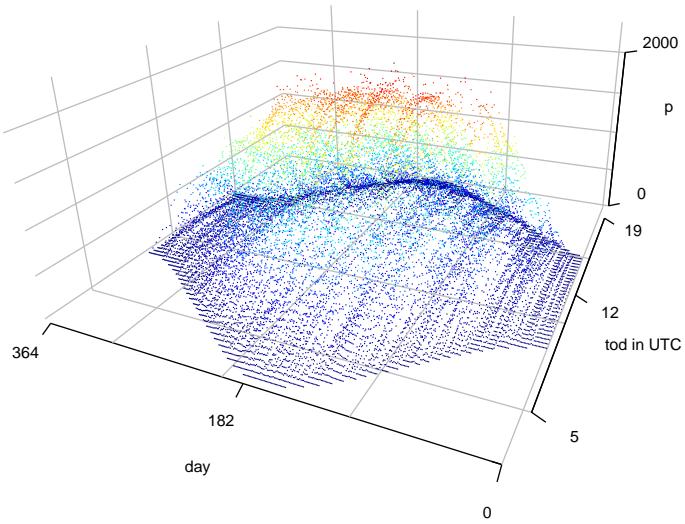


Figure 4.4: p versus t_{day} and t_{tod} . This space - spanned by t_{day} , t_{tod} and $\{P_t\}$ - is referred to as the time space.

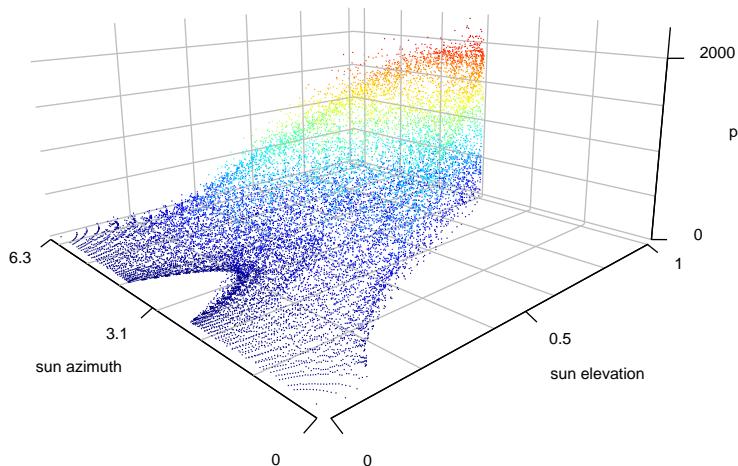


Figure 4.5: p versus ω_{se} and ω_{α} . This space - spanned by ω_{se} , ω_{α} and $\{P_t\}$ - is referred to as the sun position space.

4.1 Surface estimation

Methods for estimating the upper surface of the point clouds are many, e.g. one could find clear sky days and then simply interpolate between them etc. What is found is that the 100% quantile or a quantile close to the 100% quantile of $f_{P_t}(p_t, t_{\text{day}}, t_{\text{tod}})$ the conditional probability density function of P_t given t_{day} and t_{tod} , or the equivalent $f_{P_t}(p_t, \omega_{\text{se}}, \omega_\alpha)$, is the wanted surface. Quantile regression is used for estimating a quantile of a distribution, but since data is too sparse for each distinct time value in the dataset, the neighborhood dependency is exploited by smoothing techniques. The first method uses 2-dimensional kernel smoothing and weighted quantile regression. This method gives good results in both spaces. The second method is applied in the *time space*, where one-dimensional spline quantile regression in the t_{day} dimension gives estimates of $p_{\text{cs}}(t_{\text{time}})$ and then these estimates are smoothed in the t_{tod} dimension by smoothing splines.

The following sections describes the two methods and the statistical techniques used.

4.2 Kernel method

A 3-dimensional space (x, y, Z) , where Z is a stochastic variable and (x, y) is a coordinate set, is given. When estimating $p_{\text{cs}}(t_{\text{p},t})$ then $Z = P_t$ and either $x = t_{\text{day}}$ and $y = t_{\text{tod}}$ for the *time space* or $x = \omega_{\text{se}}$ and $y = \omega_\alpha$ for the *sun position space*. It is wanted to find an estimate of the conditional distribution $f_Z(x, y)$, so that quantile regression can be used to find the wanted quantile. This is done with 2D kernel smoothing that make weights used for weighted quantile regression. It is assumed that there is a local dependency between $f_Z(x, y)$ and $f_Z(x + d_x, y + d_y)$, which is decreasing as d_x and d_y are increasing. We have a realization of the stochastic variable

$$z = \{z_t, t = 1, \dots, N\}$$

and the corresponding coordinate set

$$x = \{x_t, t = 1, \dots, N\}, \quad y = \{y_t, t = 1, \dots, N\}.$$

4.2.1 2 dimensional kernel smoothing

A normal kernel function is used

$$w(x, x_i, h) = f_{\text{std}}\left(\frac{|x - x_i|}{h}\right)$$

where f_{std} is the standard normal probability density function. This is used in the 2D kernel function

$$k(x_i, y_i) = \frac{w(x, x_i, h_x) \cdot w(y, y_i, h_y)}{\sum_{i=1}^n w(x, x_i, h_x) \cdot w(y, y_i, h_y)}$$

where h_x and h_y are the bandwidth in respectively the x and y dimension. The bandwidths control the balance between bias and variance in the model: too low bandwidth will make variance too high and the model is overfitted, a too high bandwidth will make the model too biased and not able to estimate the curvature of the surface well enough.

4.2.2 Weighted quantile regression

It is wanted to estimate the τ_q -quantile of $f_Z(x, y)$ and this is done using the weights made by the smoothing kernel. The weighted quantile regression problem is here reduced to estimating a constant

$$\hat{Q}(\tau_q) = \hat{\beta}_{\tau_q}.$$

The loss function used is

$$\rho_{\tau_q}(\epsilon) = \begin{cases} \tau_q \epsilon & , \epsilon \geq 0 \\ (1 - \tau_q) \epsilon & , \epsilon < 0 \end{cases}$$

where $\epsilon_i = z_i - \hat{Q}(\tau_q)$. The fitting of $\hat{\beta}_{\tau_q}$ is then done by

$$\hat{\beta}_{\tau_q} = \arg \min_{\beta} \sum_{i=1}^n k(x_i, y_i) \cdot \rho_{\tau_q}(\epsilon_i).$$

where $k(x, y)$ is the kernel function. The details of the minimization can be found in [17] and is done by R-function `rq` from the package *quantreg*.

4.3 Spline method

This method makes a 2D fit by fitting in one dimension and then smoothing these fits in the second dimension. This was found obvious to do in the *time space*, since values are sampled in an equidistance grid, making it easy to estimate a conditional distribution of P_t at a given value of either t_{day} or t_{tod} . This approach works well by estimating in the t_{day} dimension and then to smooth these estimates in the t_{tod} dimension, since it is found that information of $p_{\text{cs}}(t_{\text{time}})$ is much more consistent in this dimension than in the t_{tod} dimension.

A 3-dimensional space (x, y, Z) where Z is a stochastic variable and (x, y) are a coordinate set is given. When fitting $p_{\text{cs}}(t_{\text{p},t})$ then $Z_t = P_t$ and $x = t_{\text{day}}$ and $y = t_{\text{tod}}$. We have a realization of the stochastic variable

$$z = \{z_t, t = 1, \dots, N\}$$

and the corresponding coordinate set

$$x = \{x_t, t = 1, \dots, N\}, \quad y = \{y_t, t = 1, \dots, N\}.$$

4.3.1 Non-linear quantile regression using splines

For estimating a non-linear quantile $\hat{Q}(\tau_q, x)$ of the conditional distribution $f_Z(x)$, quantile regression using splines can be applied. A non-linear model is made as

$$z_t = \beta(x) x_t + r_t.$$

The technique is to model a linear basis expansion of x_t . This is done by replacing x_t by m transformations $h_m(x_t)$, so the quantile regression model is

$$\hat{Q}(\tau_q, x_t) = \sum_{m=1}^M \hat{\beta}_{\tau_q, m} h_m(x_t) + \epsilon_t.$$

The loss function that is used to find the τ_q quantile is

$$\rho_{\tau_q}(\epsilon) = \begin{cases} \tau_q \epsilon & , \quad \epsilon \geq 0 \\ (1 - \tau_q) \epsilon & , \quad \epsilon < 0 \end{cases}$$

and $\epsilon_t = z_t - \hat{Q}(\tau_q, x_t)$. The estimation of $\hat{\beta}_{\tau_q, m}$ is done by

$$\hat{\beta}_{\tau_q, m} = \arg \min_{\beta} \sum_{t=1}^N \rho_{\tau_q}(\epsilon_t).$$

Cubic splines are used. The transformations $h_m(x_t)$ used to represent a cubic spline with k knots are

$$\begin{aligned} h_1 &= 1 , \quad h_2 = x_t , \quad h_3 = x_t^2 , \quad h_4 = x_t^3 \\ h_{4+l}(x_t) &= (x_t - \xi_l)_+^{M-1} , \quad l = 1, \dots, k. \end{aligned}$$

A cubic spline has continuous first and second derivatives at the knots. It is the lowest-order spline for which the knot-discontinuity is not visible for the eye and higher order splines are rarely more precise (p.118 in [8]).

4.3.2 Smoothing splines

Smoothing splines is used to smooth the splined quantile estimates over the second dimension. Smoothing splines is a non-linear regression method and it is an effective way to approximate the function relating two dependent stochastic variables (p. 127 in [8]). The smoothing spline used here is the function f with continuous first and second derivatives, that minimizes

$$RSS(f, \lambda) = \sum_{t=1}^N (z_t - f(y_t))^2 + \lambda \int (f''(t))^2 dt$$

where λ is a smoothing parameter that penalizes curvature in the function. The build in R-function *smooth.spline* is used to find smoothing splines.

4.4 Results

Parameter tuning and results of three estimates of $p_{\text{cs}}(t_{\text{time}})$:

1. Kernel method in the *time space*,
2. Kernel method in the *sun position space*,
3. Spline method in the *time space*,

is described in each of the following sections. The result is given as the time series

$$\hat{p}_{\text{cs}} = \{\hat{p}_{\text{cs},t}, t = 1, \dots, N\}$$

which is an estimate of $p_{\text{cs}}(t_{\text{time}})$ and

$$\tau = \left\{ \frac{p_t}{\hat{p}_{\text{cs},t}}, t = 1, \dots, N \right\} \quad (4.1)$$

is the transformed p .

Evaluating \hat{p}_{cs} is necessary to choose the best estimate and to optimize parameters of the methods. The estimates could be evaluated by the performance in predicting their outputtet τ , but this is too complicated for now, so visual inspection of the results is used. Otherwise a score function measuring the ability of a given \hat{p}_{cs} to transform p into a stationary τ , could be used. A score function is proposed as

$$\text{score}(\hat{p}_{\text{cs}}) = \int |\hat{f}_{\tau_{\text{S}}}(t_{\text{tod}_1}) - \hat{f}_{\tau_{\text{S}}}(t_{\text{tod}_2})| d\tau$$

where $\hat{f}_{\tau_S}(t_{\text{tod}})$ is an estimate of the probability density function of τ_S at a given *time of day*. Note that to be calculated and used, then $\text{score}(\hat{p}_{\text{cs}})$ have to be formulated as a discrete function. The method that minimizes $\text{score}(\hat{p}_{\text{cs}})$ function should be chosen. It is not used, but it could be used in further development of the methods for estimation of $p_{\text{cs}}(t_{\text{time}})$. Estimates of the conditional distributions for p and the three τ , is seen in Figure 4.14.

The general problem in modeling, is to find the balance between variance and bias of a model. If the model of $p_{\text{cs}}(t_{\text{time}})$ is too biased it will not model the curvature of the surface well enough, but if it has too high variance it will not exploit the information of the neighborhood enough and the surface will be too wickedly and overfitted. The parts of the surface that are difficult to estimate well, is those close to the border of positive values (i.e. in the morning and evening hours where the solar power is low) The problem of a very sparse number of clear sky days in the winter period, also makes the estimate of this part of the surface difficult. Data from more than one year, could definitely be used to improve the results of $\hat{p}_{\text{cs}}(t_{\text{time}})$ in the winter period.

For small \hat{p}_{cs} values the error of τ is obviously increasing. Therefore a lower limit α_{cut} is defined. All values of \hat{p}_{cs} where

$$\frac{\hat{p}_{\text{cs}}}{\max_i(\hat{p}_{\text{cs}})} < \alpha_{\text{cut}} \quad (4.2)$$

is removed in τ . The best level of α_{cut} for prediction models is found later in the modeling process, at page 53, but for now an appropriate value of $\alpha_{\text{cut}} = 0.2$ is used for showing the results.

4.4.1 Kernel method in the *time space*

The results of the Kernel method in the *time space*. Three input parameters are used in the method, they are

h_x Bandwidth in the t_{day} dimension. h_x controls the width of the kernel. The value used is $h_x = 35$. If this value is set higher, the level of \hat{p}_{cs} in the winter period is too low. In Figure 4.6 the plot of the kernel is shown. It is seen that h_x is much higher than h_y , i.e. the kernel cover 160 samplepoints in the t_{day} dimension, but only 5 samplepoints in the t_{tod} dimension.

h_y Bandwidth in the t_{day} dimension. h_y is much smaller relatively to h_x . $h_y = 0.2$ hours is used and if is set higher the model is too biased and especially at the border of the surface at low ω_{se} , i.e. in the morning and evening hours, the level of the \hat{p}_{cs} is too high.

τ_q The quantile estimated in quantile regression. The level of \hat{p}_{cs} in the summer period is not changed much by this parameter, but $\tau_q = 0.85$ gives a good balance in the winter period.

The result is the \hat{p}_{cs} seen in Figure 4.7. It follows data very well also in the winter period and it is just balanced to the point where it start to be varying too much. The plot in Figure 4.8 show the τ made. No extreme values are observed and level over the year is well balanced. Figure 4.9 shows the result for selected clear sky days and days with cloudy sky.

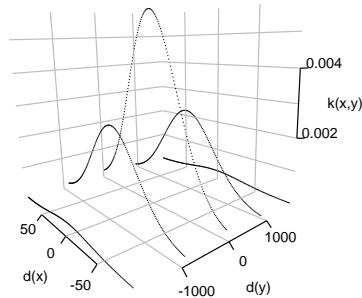


Figure 4.6: The kernel used in the time space. $d(x)$ is over t_{day} and $d(y)$ is over t_{tod} .

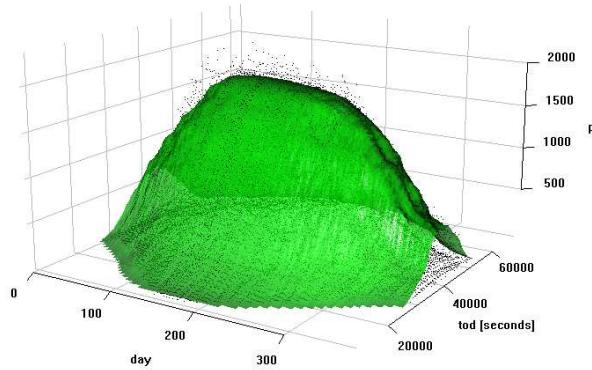


Figure 4.7: The \hat{p}_{cs} seen as the surface estimated by the Kernel method in the time space. The black points are the values of p .

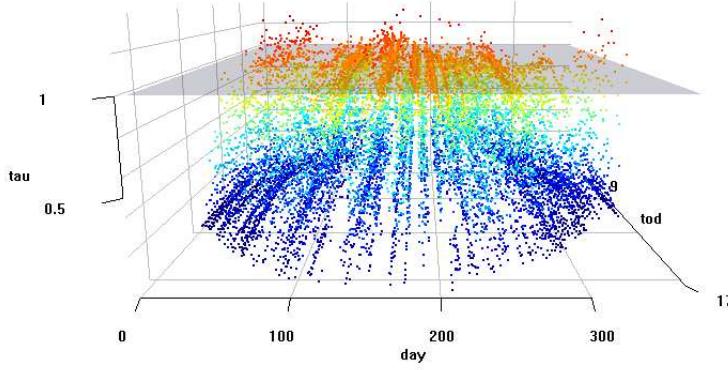


Figure 4.8: τ made with the \hat{p}_{cs} from the Kernel method in the time space.

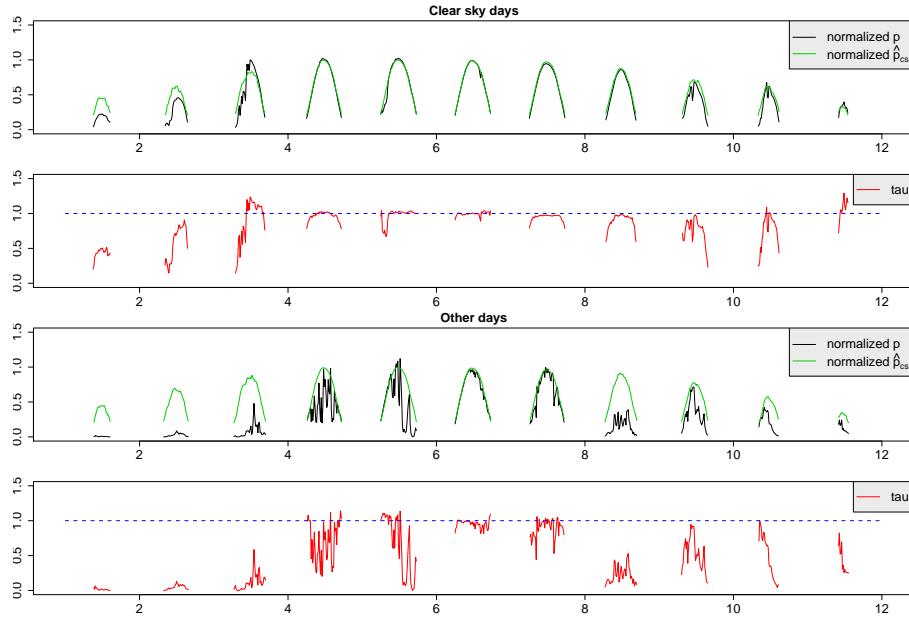


Figure 4.9: Result of the Kernel method in the time space. The upper two plots shows clear sky days evenly distributed over the year. p_{cs} is very nicely estimated in the summer period, but is getting more inaccurate towards the winter period. The two lower plots show 11 other days over the year. τ follows the dips in p nicely and is varying well between 0 and 1.

4.4.2 Kernel method in the *sun position space*

The second p_{cs} estimate is made with the Kernel method in the *sun position space*. Estimation in this space, could exploit more information from the data, because the point cloud is more dense than in the *time space*, since values in the spring and autumn can be close in terms of the position of the sun. On the other hand, other factors in the atmosphere and in the PV system could have changed the level of p_{cs} over the period. There are three parameters to the method

h_x Bandwidth in the ω_{se} dimension. $h_x = 0.1$ is the value used, higher values made the smoothing too biased. Especially in the lower range of ω_{se} this is a problem, and it can also be seen in the result that the method doesn't estimate well for this range.

h_y Bandwidth in the ω_α dimension. $h_y = \frac{\pi}{24}$ is used.

τ_q A value of $\tau_q = 0.90$ is used. This gives a good balance of the estimate between all parts of the surface.

The result is plotted in Figure 4.10 and Figure 4.11. It is seen that \hat{p}_{cs} gives a τ that is close to the wanted result also in the sparse winter period, but lacks some curvature for low ω_{se} i.e. the morning and evening hours. An improvement to the method, could be to vary the parameters as a function of ω_{se} , but this has not been tried.

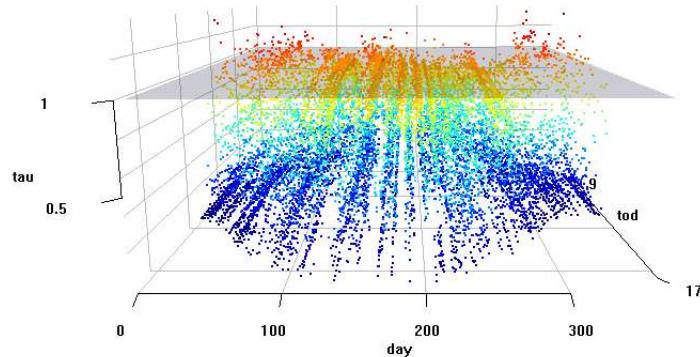


Figure 4.10: τ made by the Kernel method in the sun position space. The process is stable over the year, but is too biased in the t_{tod} dimension.

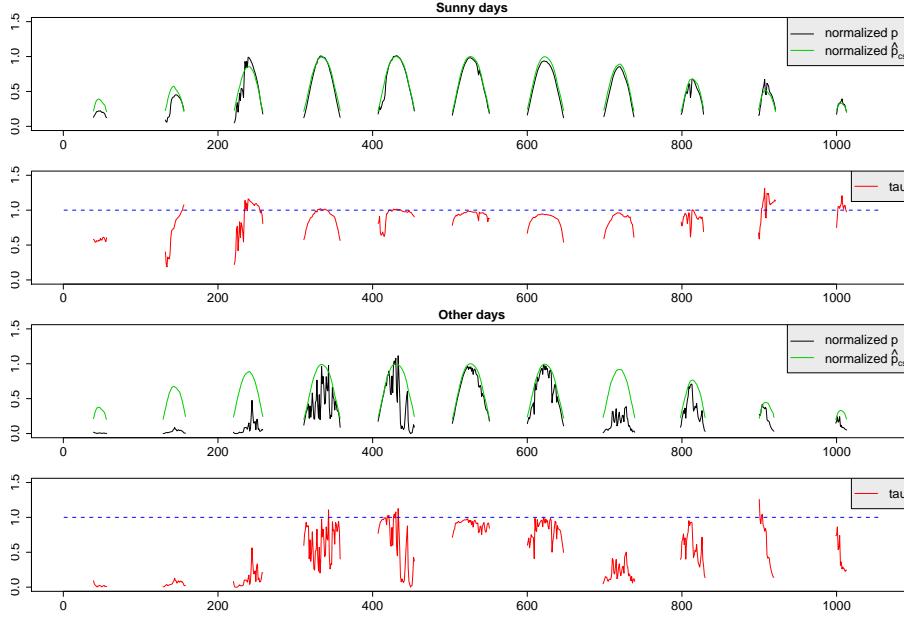


Figure 4.11: *Result of the Kernel method in the sun position space. Stable performance over the year is seen, but \hat{p}_{cs} is too biased over the day, estimating too high in the morning and evening hours.*

4.4.3 Spline method in the *time space*

The third estimate of $p_{\text{cs}}(t_{\text{time}})$ is made with the Spline method in the *time space*. The parameters used is

- τ_q The τ_q -quantile that is used when estimating in the t_{day} dimension. $\tau_q = 0.85$ is used, since it gives a good balance over the year.
- λ The parameter that penalizes curvature in the smoothing splines. $\lambda = 0.4$ gives a good balance, where the surface is not too varying, but the curvature that is needed, is just allowed.

The result is showed and commented in Figure 4.12 to 4.13.

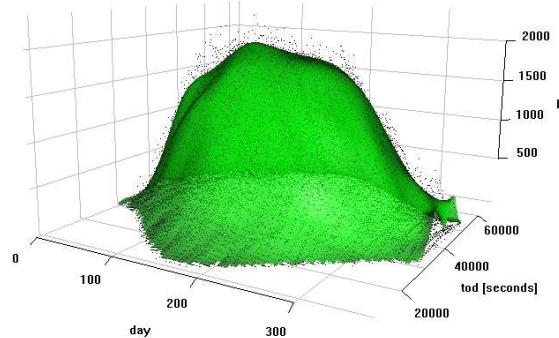


Figure 4.12: \hat{p}_{cs} made by the Spline method in the time space. Again the best result was found with parameter values that makes the surface just starting to vary. It is seen that this method goes too low in the winter period.

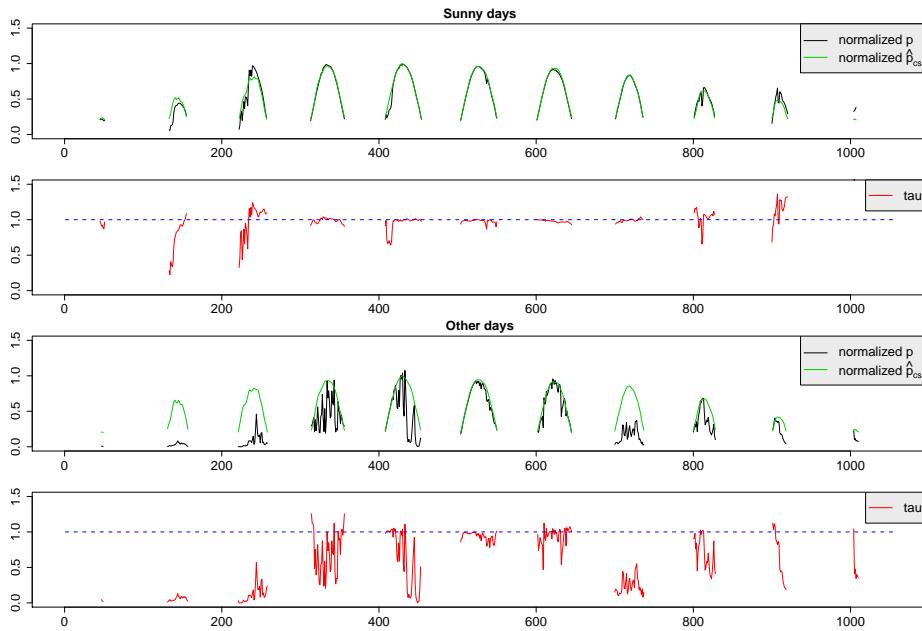


Figure 4.13: Result of the Spline method in the time space. The same days as shown for the other methods, the upper two plots are clear sky days and lower two plots are other days. The spline method seems very good in the summer period, but is too low in the winter period.

4.5 Comparison of the estimates

All three methods solves the problem well, though it immediately seems that the two estimates in the *time space* performs better, as seen in Figure 4.14, where a comparison of the conditional distribution of τ at three different values of t_{tod} - one in the morning, one at noon and one in the afternoon, is shown. They should approach each other, as the estimation of $p_{\text{cs}}(t_{\text{time}})$ is getting better. They do and seems to be a good measure of how well the methods are performing. Considering that the spline estimate has very poor performance in the winter period, the \hat{p}_{cs} made by the Kernel method in the *time space*, is used to transform p into the τ that is used in prediction models.

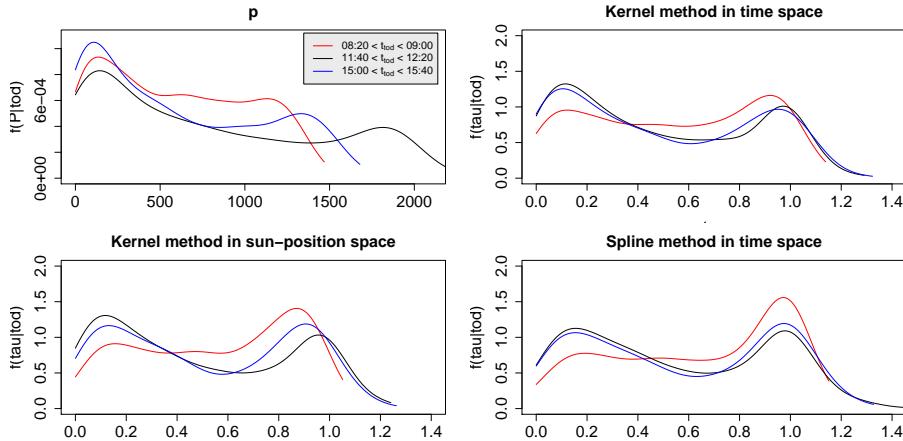


Figure 4.14: A comparison of the estimated probability distribution functions of p and τ . Three different values of t_{tod} - one in the morning, one at noon and one in the afternoon is used. The distributions are closer for τ than for p , and this indicate that $\{\tau_{S,t}\}$ more stationary than $\{P_t\}$. The time is in LST.

4.6 Conclusion

The methods used work well, though they have not been optimized and evaluated quantitatively. A question still to be investigated is if the methods are overfitted to the dataset and it could be instructive to test them on datasets from other PV systems. The methods can definitely be improved and optimized, e.g. by using more advanced smoothing techniques. Varying the parameters as a function of the input, e.g. ω_{se} , or applying a transformation of p before the estimation, could be interesting refinements to the methods. In an operational environment adaptive estimation should be applied to optimize the exploitation of information from data over longer periods of time. Data from periods over several years will certainly improve the results in the winter period. Finally it should be mentioned that $p_{cs}(t_{\text{time}})$ can also be used for interpolation of data given at low sample rates, e.g. the meteorological forecasts used in the study. The results shows that the Kernel method in the *time space* outperform the other two approaches and the \hat{p}_{cs} from this method is used to transform p into the τ used in prediction models.

CHAPTER 5

Modeling procedure

This chapter explains the procedure of building the predictions models and the theory used. The first section is about the power markets and a specification of the solar power forecasts produced is given. The next section describes how the evaluation of the prediction models is carried out. Evaluation of the prediction models is inspired by [16], where a framework for evaluating the performance of wind power prediction models is defined, and these evaluation methods can be directly applied to the solar power forecasts. When measuring the performance of the developed models a reference model is useful, especially when comparing results from different studies where the same reference model is used. A naive prediction model is developed and used as a reference model.

Finally the recursive least square (RLS) algorithm used in model estimation is derived, the level of α_{cut} (defined at page 36) to be used is decided and the chapter is ended by explaining the two types of models developed - one only using past solar power observations and another also including NWPs as input.

In the remainder of the report all time series are with hour values if nothing else is specified.

5.1 Power forecasts

The solar power forecasts will primarily be used as a tool when balancing the total power production put into the electricity grid. The balancing is done by electric power markets, where power production for a given time in the future is traded. The electric power produced in Denmark is sold on Nord Pool [19], where there are two different electric power markets are used. Elspot is a market where participants (buyers and producers) release their bids no later than 12:00 CET, for each hour the following day from 00:00 CET to 00:00 CET. The power forecasts needed to sell effectively at this market should be issued (shortly) before 12:00 CET each day and have 12-36 hours horizons. The second market is Elbas which is the regulating market where power is traded up to one hour before it is delivered. Forecasts for this market should be in 15 minute steps or finer and have very short-term horizons (less than 6 hours). They should be updated as often as possible.

To narrow the scope of this analysis, it is decided only to make hourly forecast at 12:00 UTC each day, with 1 to 36 hours horizons. This is a usable forecast type for Elspot, but it is noted that when an operational version of the forecasting method is to be made, more details have to be planned about: the update time of the NWPs, data acquisition from the PV systems, computing time and etc.

The output of the present method is forecasts of the solar power. The i 'th forecast is the time series with hour values

$$\hat{p}_i = \{\hat{p}_{i,k}, k = 1, \dots, 36\} \quad (5.1)$$

and it is made at time

$$t_{\hat{p},i} = t_{\hat{p},1} + (i - 1) \cdot 24h$$

where $t_{\hat{p},1} = 2006-01-01 12:00$. Thus the time for $\hat{p}_{i,k}$ is

$$t_{\hat{p},i,k} = t_{\hat{p},i} + k \cdot 1h - 0.5h.$$

It is noted that the k 'th horizon still is k hours. The number of forecasts made is $n = 363$, since $t_{\hat{p},364,12} = 2007-01-01 00:00$ and this is beyond the rest of the dataset.

The power forecasts are set into a matrix

$$\hat{\mathbf{P}} = \begin{pmatrix} \hat{p}_1^T \\ \vdots \\ \hat{p}_{363}^T \end{pmatrix}. \quad (5.2)$$

This is convenient, since all predictions of horizon k is in the k 'th row of this matrix, so

$$\hat{\mathbf{P}}_{1:363,k}$$

is a time series with 24 hour values containing all predictions at horizon k . This and the equivalent $\mathbf{P}_{1:363,k}$ with the measured solar power, is used when the error measures are calculated.

The prediction models are made on τ and equivalent forecasts are made of τ

$$\hat{\tau}_i = \{\hat{\tau}_{i,k}, k = 1, \dots, 36\}, \quad (5.3)$$

which are then transformed back to \hat{p}_i . These are also setup in a equivalent matrix

$$\hat{\tau}_{\text{mat}}. \quad (5.4)$$

5.2 Model evaluation

5.2.1 Error measures

The forecast are evaluated by the error measures described in this section. Most of them are found in [16]. Each error measure is calculated for each horizon, so that the quality of the prediction model is evaluated at all horizons.

The usual prediction error is

$$r_{t+k|t} = x_{t+k} - \hat{x}_{t+k|t},$$

but since only the forecasts at 12:00 is evaluated the errors used are

$$e_{t+k|t} = \mathbf{P}_{t,k} - \hat{\mathbf{P}}_{t,k},$$

and similar for τ_{mat} . When evaluating the final results then the normalized error is used

$$\epsilon_{t+k|t} = \frac{\mathbf{P}_{t,k} - \hat{\mathbf{P}}_{t,k}}{\bar{p}} \quad (5.5)$$

where

$$\bar{p} = \frac{1}{N} \sum_{t=1}^N p_t.$$

All evaluations are carried out with hour values, so $N = 8760$. The normalized error measures are calculated as the equivalent error measure defined in the following, but using $\epsilon_{t+k|t}$ instead of $e_{t+k|t}$ and are pre-scripted with an N, e.g. *NRMSE* is Normalized *RMSE*.

The Root Mean Square Error of the k 'th horizon is

$$RMSE_k = \left(\frac{1}{n} \sum_{t=1}^n e_{t+k|t}^2 \right)^{\frac{1}{2}}.$$

The $RMSE_k$ is used as the main criteria for the performance of the models.

Since samples below α_{cut} have been removed, resulting in missing forecast values, a measure that evaluates how complete the forecast are for horizons k , it is defined as

$$Completeness_k = \frac{n_{\text{forecasts}} - n_k}{n_{\text{forecasts}}}$$

where $n_{\text{forecasts}}$ is the maximum number of forecast that can be made in the given context, thus $n_{\text{forecasts}} = 363$ is used for $\hat{\mathbf{P}}$, and n_k are the number of forecasts at horizon k made by the model.

Further evaluation of a model includes the

$$BIAS_k = \frac{1}{N} \sum_{t=1}^N \epsilon_{t+k|t},$$

which is the systematic error of the model.

$BIAS_k$ is associated with the first moment of the prediction error. $RMSE_k$ is related to the second moment and it penalizes large errors relatively higher [16].

When comparing the performance of two models the improvement for horizon k is used

$$I_{EC,k} = 100 \cdot \frac{EC_{\text{ref},k} - EC_k}{EC_{\text{ref},k}} (\%)$$

where EC is the considered evaluation criteria (i.e. error measure) used. The completeness of one model relative to the other, is defined as

$$rCompleteness_k = \frac{n_k}{n_{\text{ref},k}}$$

where n_k are the number of forecasts at horizon k made by the new model and $n_{\text{ref},k}$ is the number of forecasts made by a reference model at horizon k .

5.2.2 Evaluation of a model

When identifying the best prediction model, the model achieving the best performance predicting τ is wanted. The main evaluation is made by $RMSE_k$ and

to ensure that the result is not too sparse the $Completeness_k$ is always checked. An example of these calculated on the output of a model is found in Figure 5.2. Inspired by [16] further exploration of the performance of a model is done by evaluating $BIAS_k$ and MAE_k and plotting the prediction error versus the predicted value. Histograms estimating the distribution of the prediction error are made and they are used to find weak points of a model. The cumulated errors can be used to find periods and conditions, where the model performance is poor.

5.2.3 Comparison of two models

An important point when comparing the result of two models, is that only values that are predicted by both models are used in the calculation of the error measures. The main comparison of two models is done by the $I_{NRMSE,k}$. Models including more explanatory values will often have a lower $Completeness_k$ and therefore the $rCompleteness_k$ is also used where it is found relevant.

5.3 Reference model

A model that sets a reference level for comparing performance of advanced models is needed. A reference model can also be used to compare model performance between studies that use similar reference models. A detailed development of a reference model has not been found in other studies, and the wind power evaluation in [16] defines a reference model that is not directly applicable for solar power forecasts. Previous studies use different persistence reference models. A reference model is developed here, where the best naive predictor for each distinct k horizon is chosen. Three different models are tried. The first is referred to as the Diurnal-persistence model

$$\hat{p}_{t+k|t} = \begin{cases} p_{t+k-24} & \text{for } 1 \leq k \leq 24 \\ p_{t+k-48} & \text{for } 25 \leq k \leq 36 \end{cases}$$

i.e. the last value observed at the same *time of day* is used for prediction. This reference model is used in other studies e.g. [14]. The second is the Persistence model

$$\hat{p}_{t+k|t} = p_t.$$

and the third is the Global-mean model, where the mean over the j 'th sample of p each day is used, as

$$\bar{p}_{\text{tod},j} = \frac{1}{363} \sum_{i=0}^{363} p_{(j+i \cdot 24)}$$

so the model is

$$\hat{p}_{t+k|t} = \bar{p}_{\text{tod},((t+k) \bmod 24)}.$$

The $RMSE_k$ for the three naive prediction models at all horizons is showed in Figure 5.1. The model having the lowest $RMSE_k$ is chosen for that horizon. The Reference model is then

$$\hat{p}_{\text{ref},(t+k|t)} = \begin{cases} p_t & \text{for } 1 \leq k \leq 2 \\ p_{(t+k-24)} & \text{for } 3 \leq k \leq 24 \\ p_{(t+k-48)} & \text{for } 25 \leq k \leq 36 \end{cases}$$

When using the Reference model for comparison with prediction models on τ , it is transformed into

$$\hat{\tau}_{\text{ref},(t+k|t)} = \frac{\hat{p}_{\text{ref},(t+k|t)}}{\hat{p}_{\text{cs},t+k}}$$

and then the $RMSE_k$ and $Completeness_k$ is calculated as shown in Figure 5.2.

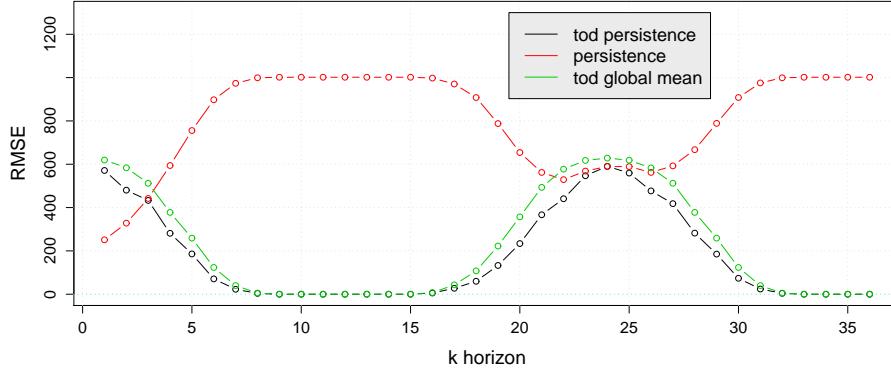


Figure 5.1: $RMSE_k$ of the different naive prediction models. It is clearly seen that the two persistence models performs better than the Global-mean and that Diurnal-persistence is best except for the first two hours. The Diurnal-persistence model has an unexpected dip at $k = 2, 22$ and 26 . The model implementation has been checked thoroughly and is found to be correct.

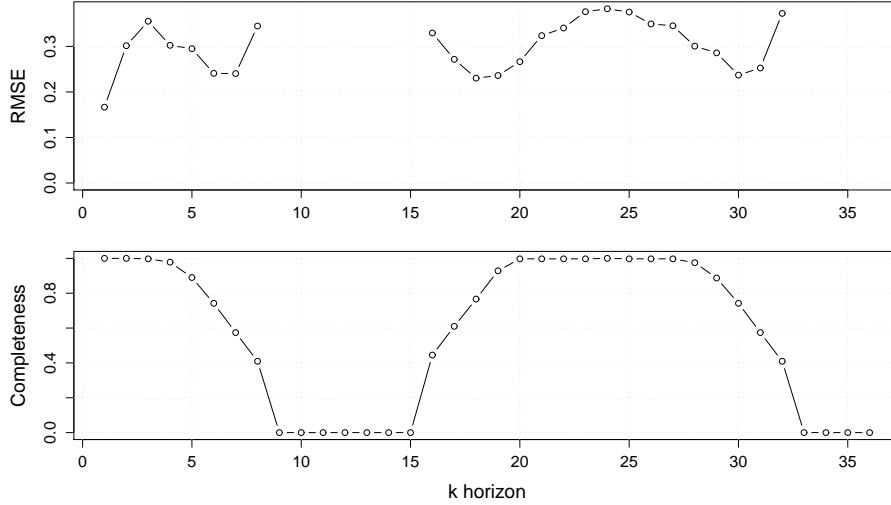


Figure 5.2: Evaluation of the Reference model when transformed into τ . It is seen that because of poor estimation of p_{cs} toward dawn and dusk, the $RMSE_k$ is increasing in these horizons. The $Completeness_k$ falls to 0 at nighttime.

5.4 k -step recursive least squares with forgetting

When making an ARX time series model as

$$Y_{t+k|t} = m + a_1 Y_t + \dots + a_p Y_{t-(p-1)k} + b_1 \hat{Z}_{t+k|t} + e_{t+k}$$

where p is the number of parameters, then methods for estimating the optimal coefficients for prediction is needed. These methods have been developed and explored to rich extend and here just an explanation of the methods used, is given. The model above is used as an example and it can obviously be extended with more input parameters. A least squares fit minimizes the $RMSE_k$ and since the model is changing over time, adaptive estimation is a necessity to make the optimal model. More details of time series modeling used, is found in [15]. The adaptive estimation is done by k -step RLS with exponential forgetting, as explained in the following.

The regressor vector at t is

$$\mathbf{X}_t^T = (1, Y_t, \dots, Y_{t-(p-1)k}, \hat{Z}_{t+k|t})$$

and the parameter vector

$$\theta_t^T = (m, a_1, \dots, a_p, b_1),$$

so that the model can be written as

$$Y_t = \mathbf{X}_t^T \theta_t + e_t.$$

The loss function at t

$$S_t(\theta_t) = \sum_{s=1}^t \lambda^{t-s} (Y_s - \mathbf{X}_s^T \theta_t)^2$$

makes weighted least squares with exponential forgetting and the algorithm used estimates the parameters so that

$$\hat{\theta}_t = \arg \min_{\theta} S_t(\theta).$$

The solution to the normal equation at t (details in [15]) is

$$\hat{\theta}_t = \mathbf{R}_t^{-1} \mathbf{h}_t$$

where

$$\mathbf{R}_t = \sum_{s=1}^t \lambda^{t-s} \mathbf{X}_s \mathbf{X}_s^T, \quad \mathbf{h}_t = \sum_{s=1}^t \lambda^{t-s} \mathbf{X}_s Y_s.$$

The k -step RLS-algorithm with exponential forgetting is then derived as

$$\begin{aligned}\mathbf{R}_t &= \lambda \mathbf{R}_{t-1} + \mathbf{X}_{t-k} \mathbf{X}_{t-k}^T \\ \hat{\theta}_t &= \hat{\theta}_{t-1} + \mathbf{R}_t^{-1} \mathbf{X}_{t-k} (Y_t - \mathbf{X}_{t-k}^T \hat{\theta}_{t-1})\end{aligned}$$

and the k -step prediction at t is

$$\hat{Y}_{t+k} = \mathbf{X}_t^T \hat{\theta}_t.$$

The starting values used is $\mathbf{R}_0 = (1000 \mathbf{I})^{-1}$ and $\theta_0^T = (0, \dots, 0)$, where \mathbf{I} is the identity matrix.

Making this run properly with data containing missing values, updating of the parameter vector can only be done when a complete set for the iteration exist. This implies that at t , both

$$\mathbf{X}_{t-k} \text{ and } Y_t$$

must be complete with no missing values to make an update. If the set is not complete then the previous values are kept

$$\begin{aligned}\mathbf{R}_t &= \mathbf{R}_{t-1} \\ \theta_t &= \theta_{t-1}.\end{aligned}$$

Similarly a is prediction only made at t if \mathbf{X}_t is complete.

It is noted that the optimal value of λ needs to be found, and this is done for each model as described in Section 6.1 on page 58. Furthermore a burn-in period is needed where predictions are not evaluated. It should have a sufficient length, so that enough samples are observed to settle the parameter estimates. By experience with the current data covering one year and containing missing values, this is set to $\frac{1}{4}n_{\text{upd}}$ where n_{upd} is the number of θ updates performed. If more than one year of data was available, this period could just be of constant length.

5.5 Finding the best α_{cut} limit

Because the Clear Sky model is performing poor toward dawn and dusk, where small values are observed, samples are removed below the α_{cut} limit, as defined and explained on page 36. An appropriately level of α_{cut} is found by evaluating the result of a simple k -step seasonal AR model

$$\hat{\tau}_{(t+k|t)} = m + a_1 \tau_t + e_{(t+k)}.$$

The results for different values of α_{cut} is shown in Figure 5.3. It is noted that the maximum of Completeness_k for the AR models starts at 0.75, since predictions

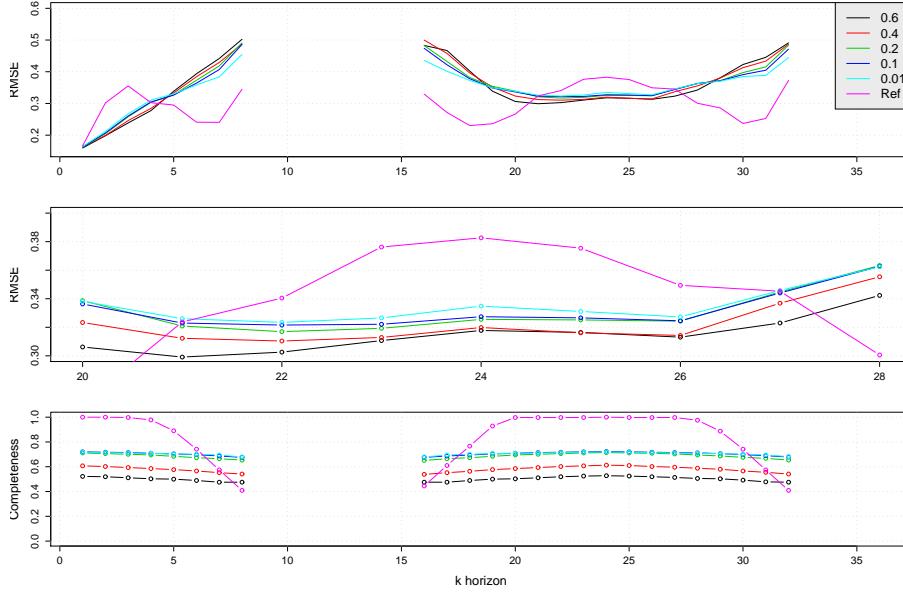


Figure 5.3: Result of a simple AR(1) k -step seasonal model for different values of α_{cut} and the Reference model. The increase in $RMSE_k$ for the AR models toward dawn and dusk, is most likely caused by poor performance of \hat{p}_{cs} . The opposite shape for the Reference model where lower $RMSE_k$ is seen throughout the day, is because the error of \hat{p}_{cs} is not influencing the predictions there. Though at the last couple of horizons toward dawn and dusk quantization errors in \hat{p}_{cs} influence the transformed values.

in the burn-in period are removed. The performance regarding $Completeness_k$ start to differ at $\alpha_{\text{cut}} = 0.2$ where almost maximum $Completeness_k$ is still found, and $RMSE_k$ is better than for $\alpha_{\text{cut}} < 0.2$. Investigating Figure 5.4(a) shows a negative I_k down to -4% for $\alpha_{\text{cut}} = 0.1$ vs. $\alpha_{\text{cut}} = 0.6$ turning to a positive improvement toward dawn and dusk. Figure 5.4(b) reveals very small $I_{RMSE,k}$ differences for $\alpha_{\text{cut}} = 0.1$ vs. $\alpha_{\text{cut}} = 0.2$. An important point when reviewing the evaluations is seen here. The error measures in Figure 5.4 which are only calculated with the values predicted by both models, show at e.g. $k = 20$ equal performance, but in Figure 5.3 where all predictions by each model are included in the error measures, the result at $k = 20$ is different. Hence the values removed at $k = 20$ for $\alpha_{\text{cut}} = 0.6$ but included for $\alpha_{\text{cut}} = 0.1$, are values where the $RMSE_k$ is higher than for those included for both α_{cut} levels. Around noon $I_{RMSE,k}$ seems to be more correct in Figure 5.3, but the $Completeness_k$ is still lower for $\alpha_{\text{cut}} = 0.6$. The conclusion is that $\alpha_{\text{cut}} = 0.2$ is a reasonably level, since a high performance is achieved while keeping close to maximum $Completeness_k$ for all horizons.

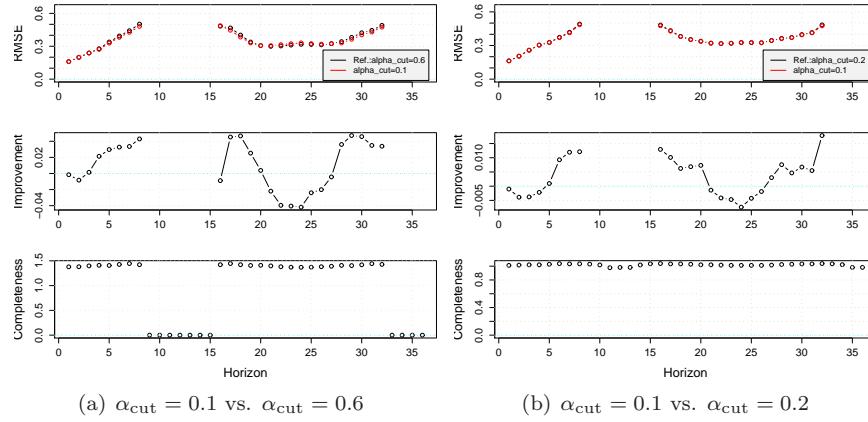


Figure 5.4: (a): Positive $I_{RMSE,k}$ are observed towards nighttime, but around noon $\alpha_{cut} = 0.6$ is better, but the $Completeness_k$ is higher for $\alpha_{cut} = 0.1$ at all horizons. (b): No considerable difference is observed between $\alpha_{cut} = 0.2$ and $\alpha_{cut} = 0.1$.

5.6 Modeling procedure

Finding the best prediction model is considered in two scenarios:

- Models with no external input, only past observations of the solar power is used as input, i.e. it is an autoregressive (AR) model. It is of course interesting to find out how well the solar power can be predicted without no external input, since it is not always possible to acquire external information.
- Models using NWP as input. By using NWPs the results should be improved and the model using both types of input is expected to be the best performing model found. This is an autoregressive with exogenous input (ARX) model.

The procedure of finding the best model is inspired by [15], where the starting point is to estimate the ACF of the time series and decide on a suitable model. This model is then fitted, with RLS where the best λ is chosen, and the ACF of the residuals is investigated. The model is expanded with only one more input variable at a time - or one term is removed - until no more improvement is gained. To be sure that no models with better performance are missed, comparisons of the selected model to all other smaller models have been carried out, and it was found that all smaller models have decreased performance.

When using NWPs as input the model with no solar power input is investigated first, and then these are added one by one. The performance of the final selected

model in each scenario is then further explored.

CHAPTER 6

Autoregressive models

A reasonable starting point for developing a prediction model solely based on the past solar power observations, is by using linear autoregressive (AR) models. The models are made on the cloud cover time series with hourly values

$$\tau = \{\tau_t, t = 1, \dots, 8760\}$$

which is defined in Chapter 4. The model parameters are adaptively estimated with the RLS method defined in Section 5.4. The models are fitted for each horizon and it is thus possible to use different parameters and models for different horizons.

6.1 Model identification

Identifying the best model is done as described in Section 5.6. The autocorrelation function (ACF) of τ is plotted in Figure 6.1. Clearly there is an exponentially decreasing trend in the first couple of lags indicating an AR(1) component, and a similar trends is observed with a seasonal lag of 24. Hence the first model evaluated is

$$\hat{\tau}_{t+k|t} = m + a_1 \tau_t + b_1 \tau_{(t-24+k \bmod 24)}.$$

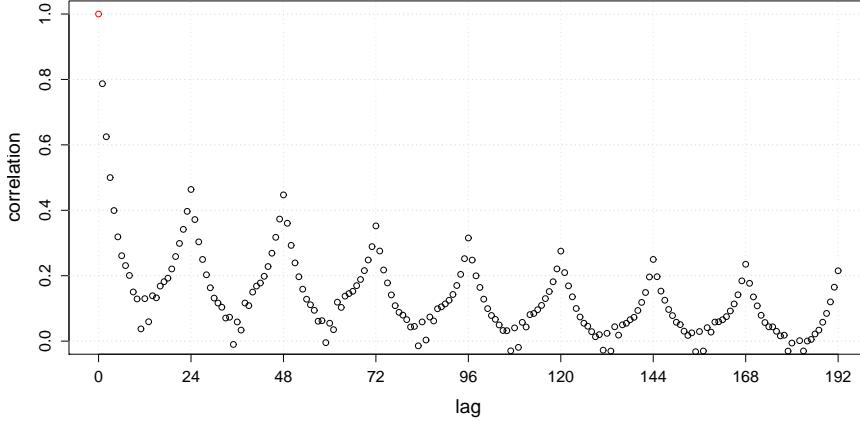


Figure 6.1: *ACF of τ . An exponentially decreasing trend is seen for the first couple of lags and for seasonal lags of 24, 48, ...*

The $\tau_{(t-24+k \bmod 24)}$ term is the 24 seasonal lag for $t + k$, and the modulus is needed since for $k > 24$ then using $t - 24 + k$ would result in a future value of τ being used and this is of course not possible. To name the seasonal part of the model a discrete function

$$k_s = l(k, s) = s \cdot \left\lfloor \frac{k}{s} + 1 \right\rfloor$$

gives the seasonal lags used in the AR model for k . Thus the above model is named $AR_{k,k_{24}}$.

The model is fitted using the RLS method and this implies finding an optimal value of the forgetting factor λ . This is done by simply fitting with $\lambda = 0.95, 0.955, \dots, 1$, calculating $RMSE_k$ for each lambda and horizon, inspecting the result visually and then deciding the best value. The plot is shown in Figure 6.2 where $RMSE_k$ for $k = 1, 3, \dots, 36$ are shown. All other k horizons have also been inspected. A λ value of 0.995 is found to be the best balance for all horizons. An optimized λ value is chosen in the same manual way for all other models used. The λ value used is noted when they are introduced.

The ACF of the residuals of $AR_{k,k_{24}}$ is plotted in Figure 6.3. $k = 1, \dots, 3$ shows that information is effectively removed for a couple of lags, but as shown by the red dots lags smaller than $k - 1$ can not be used and therefore they have the same trends as in the ACF of τ . The step from $k = 23$ to 24 shows that the peak at lag 48 is removed, since it is included in the model, and from $k = 24$ to 25 it is seen that the peak at lag 24 starts to increase immediately. It is observed that for $k < 24$ there is a peak at lag 48, but no 48 lag component in the model

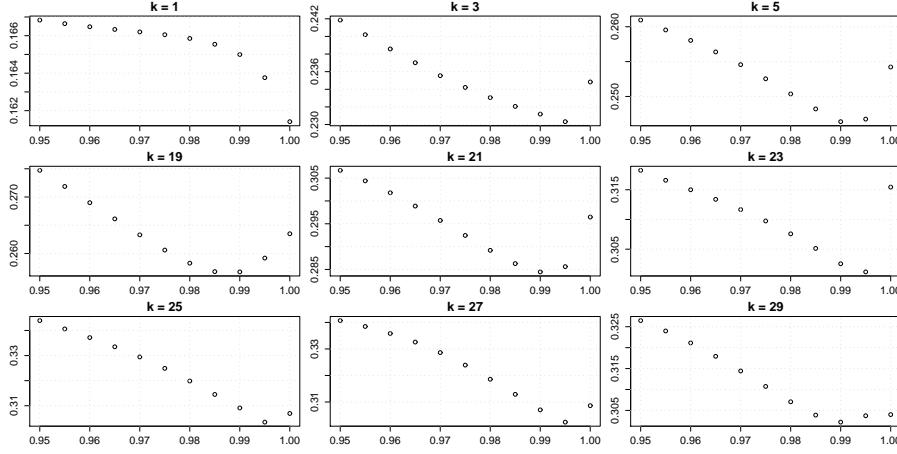


Figure 6.2: $RMSE_k$ for $\lambda = 0.95, \dots, 1$ for model $AR_{k,24}$. Only plots of odd horizons with a result is included. $\lambda = 0.995$ is chosen as the best balance for all horizons.

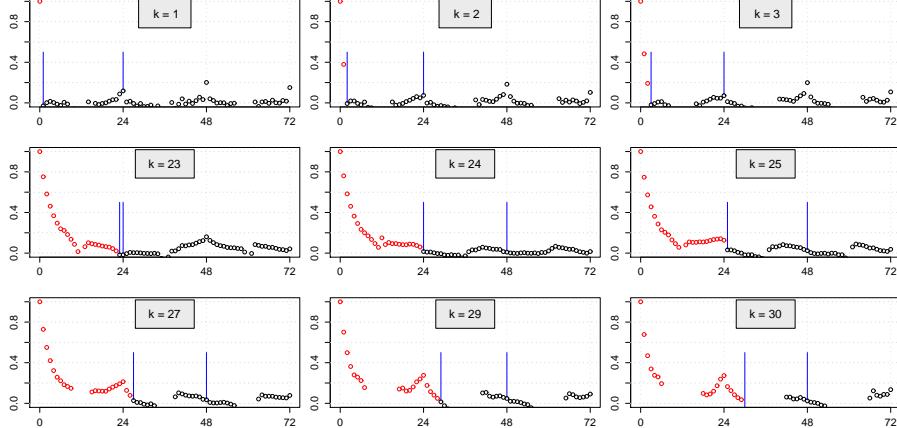


Figure 6.3: ACF of the residuals from $AR_{k,k_{24}}$ for selected horizons. The blue lines indicate the lags used in the model. The red dots are lags smaller than $k - 1$ that can not be used for the horizon k .

and therefore this component is added to the model. The conclusion is to add a seasonal component of lag 48, so the next model evaluated is

$$\hat{\tau}_{t+k|t} = m + a_1 \tau_t + b_1 \tau_{(t-24+k \bmod 24)} + b_2 \tau_{(t-48+k \bmod 24)}.$$

and this is denoted by $AR_{k,k_{24},k_{48}}$. Optimizing for λ showed same pattern as seen in Figure 6.3 and $\lambda = 0.995$ is chosen again. The ACF of the residuals is showed in Figure 6.4 for the same horizons as previous. For $k = 1, 2, 3$ the peak lag 48 is not decreased much compared to Figure 6.3. The high ACF at lag 24,48,72 is believed to come from non-stationarity in τ and thus it is found that all dependency of time of day has not been removed by the Clear Sky model. For $k = 23$ it seems as if there is an improvement using lag 48. For $k = 24, 25, 27, 28, 29$ no difference is observed. A comparison between $AR_{k,k_{24},k_{48}}$ and $AR_{k,k_{24}}$ as the reference, the result is showed in Figure 6.5. The results for $k = 20, \dots, 23$ indicate an improvement - as expected from the ACF - but since no overall improvement is observed, it is decided not to include the 48 lag seasonal term.

A model

$$\hat{\tau}_{t+k|t} = m + a_1 \tau_t + b_1 \tau_{(t-24+k)} + b_2 \tau_{(t-48+k)}.$$

where each seasonal terms are dropped one by one when it is not possible to use them, has also been compared to $AR_{k,k_{24}}$. The improvement result is very much alike the result in Figure 6.5 for $k = 1, \dots, 23$ and the improvements are equal for $k = 24, \dots, 30$. Since it is judged that the performance of these two model are equivalent for $k = 1, \dots, 6$, and $AR_{k,k_{24}}$ is the smaller of the two models, $AR_{k,k_{24}}$ is selected.

Though nothing in the ACF found yet clearly indicates that other lags could be used

$$\hat{\tau}_{t+k|t} = m + a_1 \tau_t + a_2 \tau_{t-1} + b_1 \tau_{(t-24+k \bmod 24)}.$$

is evaluated and it showed $I_{RMSE,k}$ between -0.01 and 0.01 with no clear pattern. All smaller models have been evaluated and compared to $AR_{k,k_{24}}$ - also models without an intercept term - and it is found that they have a decreased overall performance.

Thus it is concluded that the best simple AR model of τ is $AR_{k,k_{24}}$ without overfitting the model to the data. It is also found that a more thorough analysis could conclude on different models for intervals of k - e.g. including a seasonal lag 48 for $k = 20, \dots, 23$ - but this is left for further work. It is also noted that since τ is not perfectly stationary then a better estimation of \hat{p}_{cs} could give a different result.

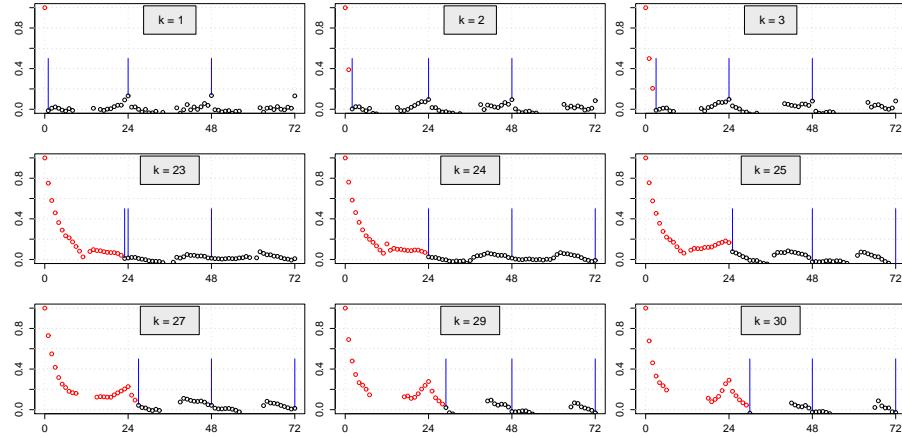


Figure 6.4: *ACF of residuals from $AR_{k,k_{24},k_{48}}$. The blue lines indicate the lags used in the model. The red dots are lags smaller than $k - 1$ that can not be used for the horizon k .*

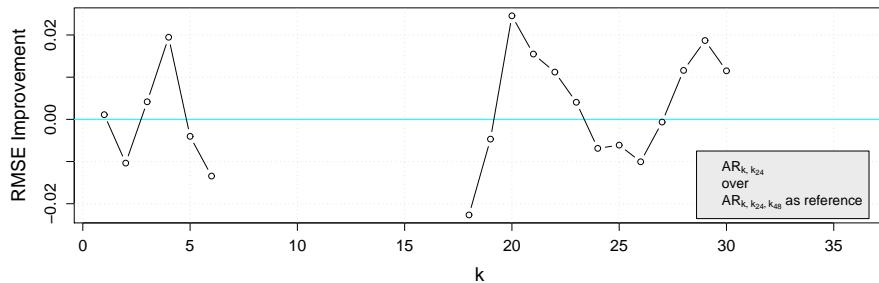


Figure 6.5: *The $I_{RMSE,k}$ of $AR_{k,k_{24},k_{48}}$ over $AR_{k,k_{24}}$ as reference. It is assessed that no improvement is achieved.*

6.2 Evaluation of the selected AR model

The selected model

$$\hat{\tau}_{t+k|t} = m + a_1 \tau_t + b_1 \tau_{(t-24+k \bmod 24)}.$$

denoted by $AR_{k,k_{24}}$, is evaluated in this section. All result are concerning τ and when the Reference model is used, it is transformed by \hat{p}_{cs} . First the model is compared to the Reference model by calculating $RMSE_k$ for each horizon, this is shown in the upper plot of Figure 6.6. The result show a very distinct improvement, except for the first lag, thus as for wind power [16] it seems as if the naive predictor is hard to beat for very short horizons less than 1 hour. The middle plot show $BIAS_k$ for each horizon and it is observed that it is relatively high for the last horizon toward dawn and dusk. It is considered to be an effect of the poor estimation of p_{cs} toward dawn and dusk. A distinct pattern for $k = 20, \dots, 29$ - the last part is also found for $k = 1, \dots, 5$ - is found, but if the magnitude of the $BIAS_k$ has a considerable effect is not concluded. The lower plot shows the $Completeness_k$ for each horizon. It is decreasing towards the nighttime and this is a directly consequence of the removal of values in τ due to the α_{cut} limit as explained on page 36.

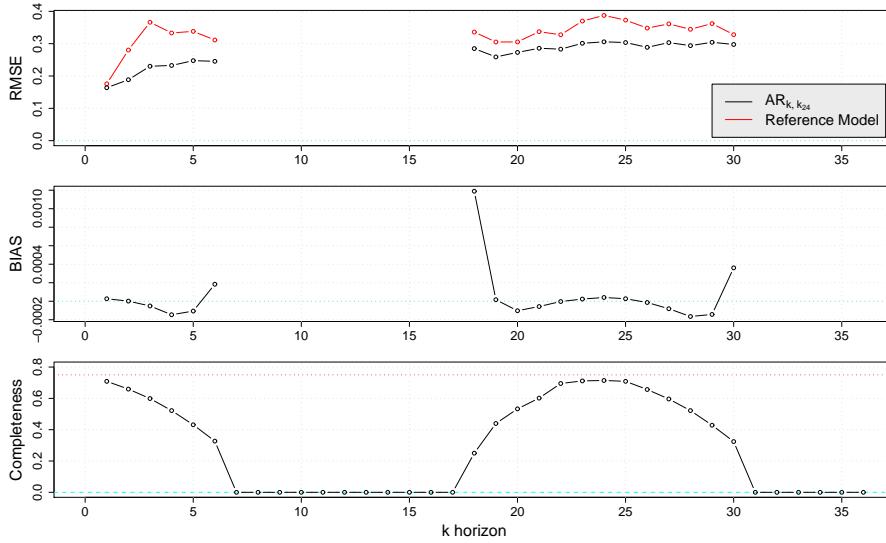


Figure 6.6: Upper: The $RMSE_k$ of $AR_{k,k_{24}}$ and the Reference model. Middle: $BIAS_k$ of $AR_{k,k_{24}}$ for each horizon. Lower: The $Completeness_k$. Note that the $Completeness_k$ only scale up to 0.75. Due to the length of the burn-in period the first 25% of the predictions are removed.

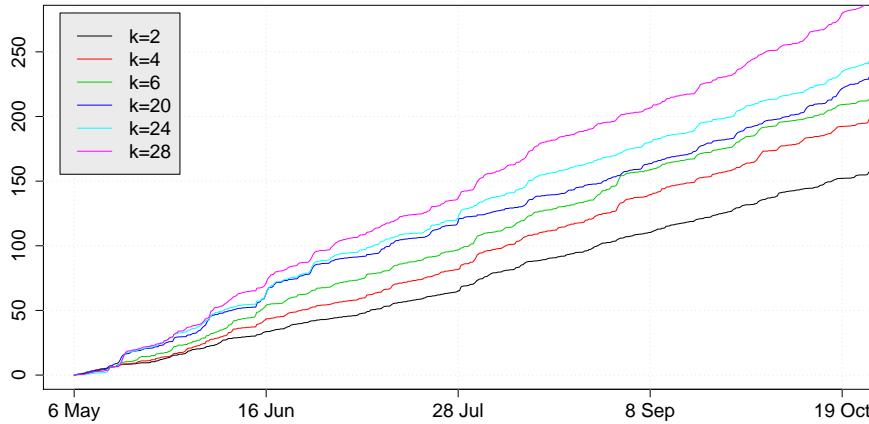


Figure 6.7: The cumulated squared prediction errors over the summer period for six selected horizons. Almost a half year is shown. There is no indication of seasonal changes in the global trend of the errors, but local changes are clearly seen.

A model with low $Completeness_k$ is of course a worthless prediction model, but in this stage of explorative modeling where the prospect of an operational model still is some way ahead, the observed $Completeness_k$ is found acceptable.

To investigate the performance over time the cumulated squared prediction errors are plotted in Figure 6.7. The overall trend in each of the six lines is found to be without any change over the period, but local changes in the trend over periods of several days are clearly seen. These periods should certainly be investigated more closely, to find what causes the high errors. For $k = 20, 24, 28$ the periods of local changes occurs nearly at the same time, but for $k = 4, 6$ the changes occurs in other periods. It leads to think that different types of weather causes uncertainties for different horizons. A similar plot over several years, would maybe show some difference between the winter and summer period.

The distribution of the prediction errors are plotted in Figure 6.8 for four selected horizons. First of all it is seen that the errors are centered well around 0, though not perfectly symmetric. $k = 18$ is, as expected from the $BIAS_k$ result shown in Figure 6.6, most skewed. The sharpness is decreasing as the horizon increases and that the relative number of times errors are above a certain limit e.g. 0.5, is increasing as the horizon increases. The histograms of all values at a given horizon is found not to reveal any weak points of the $AR_{k,k_{24}}$ prediction models. Distributions of prediction errors conditioned on various variables, e.g. $\hat{\tau}$, would be able to reveal more details of the model performance, should therefore be explored in further work. Furthermore qqplots have been checked for each horizon, without finding any noticeable behavior of the prediction errors.

The prediction errors versus the predicted value $\hat{\tau}$ is shown in Figure 6.9. This gives an indication of how the distribution of the errors are conditioned on $\hat{\tau}$. Furthermore the range of $\hat{\tau}$ is divided into 10 equally sized bins, and the BIAS is calculated for each and showed by a red line. Note that the scale of the BIAS is in the right side of the plot. The errors seems to be quite well balanced around 0 for all levels of $\hat{\tau}$ and no clear pattern is seen in the BIAS for the different bins. Naturally the errors are within a band, since the values of $\hat{\tau}$ are all in the range of 0 and approximately 1. Hence say that the maximum value of $\hat{\tau}$ really was 1, then

$$\hat{\tau} \rightarrow 1 \Rightarrow P(e < 0) \rightarrow 1$$

where P is the probability. This can not be avoided and it does not indicate any shortcomings of the model. It is quite apparent, that the sharpness of the distribution of the errors is lower for $\hat{\tau}$ in the middle of its range, than for $\hat{\tau}$ values toward its lower and upper limits. This information should be used when the uncertainties of the forecasts are modeled.

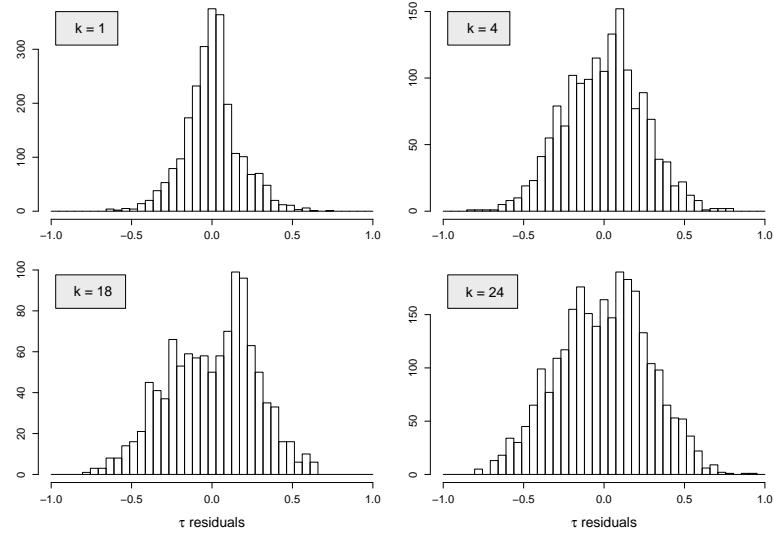


Figure 6.8: Distribution of the prediction errors from $AR_{k,k_{24}}$ for four selected horizons.

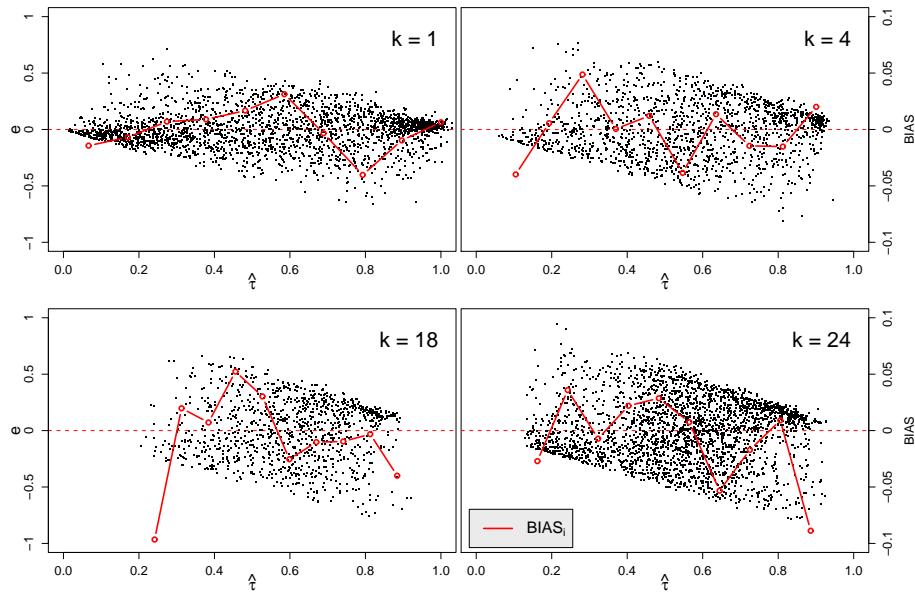


Figure 6.9: The prediction errors as a function of $\hat{\tau}$. The red line is the BIAS for 10 bins. Note the different scales.



Figure 6.10: The coefficient estimates of $AR_{k,k_{24}}$ seen over time for four selected horizons. The burn-in period is marked with a grayed background. The optimal $\lambda = 0.995$ is used. Clearly changes of the estimates are seen.

The RLS estimates of the coefficients in $AR_{k,k_{24}}$ over time is shown in Figure 6.10. Clearly changes over short periods of a few days are observed. Often when the level of one coefficient is decreasing another is increasing, and it would be obvious to think that these periods are also the periods of high errors observed in the accumulated errors shown in Figure 6.7. Indications of annual trends with slow changes of the levels is also noted, and it is found likely that data from more years would show annual periodicity.

CHAPTER 7

Autoregressive models with exogenous input

This chapter describes a prediction model build using both past observations of solar power and NWPs as input. Such a model is an Only NWPs of Global irradiance is used. Previous studies as [13] show that the solar power is also dependent on temperature, wind speed and wind direction, but the aspects have to be explored in future studies of this data. The second section identifies the best model with NWP input and finally this model is evaluated.

7.1 Transforming global irradiance to solar power

The function between global irradiation at the location of the PV system and the solar power from the system is needed, so that the NWPs of global irradiance can be transformed into predictions of solar power. It is denoted by

$$power(G).$$

Observations of irradiance and solar power P_{module} (see section 3.4) is used to explore how to model this function. Two irradiance signals are measured: global irradiance G_{obs} and irradiance incident to the surface of the PV module $I_{\text{surf,obs}}$.

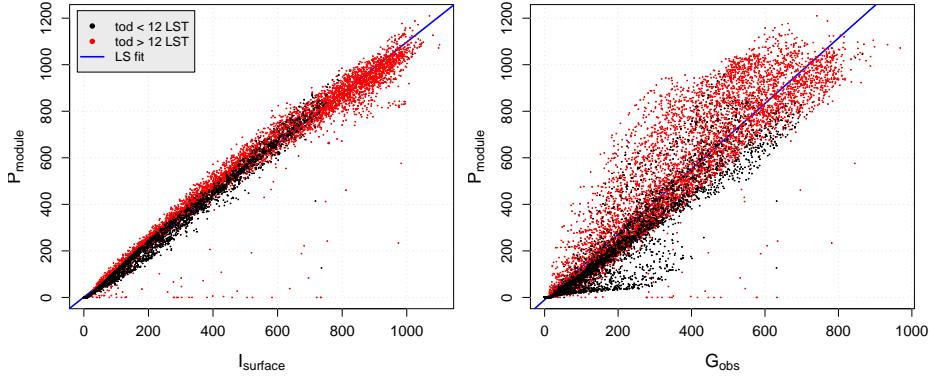


Figure 7.1: Left: solar power P versus the irradiance incident to the surface of the PV module I_{surface} . It is found reasonable to assume that the function is linear. Right: A clear indication of linear dependency is seen, but with a tendency of a higher slope for $t_{\text{tod}} > 12 \text{ LST}$.

By studying P_{module} versus $I_{\text{surf,obs}}$, plotted in the left plot of Figure 7.1, it is found that $\text{power}(G)$ is well approximated by a linear function. It is noted that irradiance observations are done with a sensor, which is simply a small PV element compensated for temperature dependency. Hence the variation around the LS fit in the left plot could very well be caused by variation in the temperature. The right plot show P_{module} versus G_{obs} . A few outliers where the solar power $P_{\text{module}} = 0$ and $G_{\text{obs}} > 200$ is assessed to be faulty observations. A deviation from the main trend for values in the morning hours is found interesting, but is not investigated further. The main effect noticed is a clear indication of a different distribution for values where $t_{\text{tod}} > 12$ compared to values where $t_{\text{tod}} < 12 \text{ LST}$. This effect is investigated further in Figure 7.2, where it is noted that all values where $G_{\text{obs}} < 20$ have been removed. A clear indication of a dependency between $\frac{P_{\text{module}}}{G_{\text{obs}}}$ both of the *time of day* and of the *time of year* is observed. These two dependencies are also found in the NWPs of global irradiance and the following describes how $\text{power}(G)$ is modeled.

The NWPs used is explained and defined in Section 3.3. The forecasts of global irradiance is denoted by

$$\hat{g}_i = \{\hat{g}_{i,k}, k = 1, \dots, 16\}.$$

This \hat{g}_i is a time series of 3 hour values. The time series of solar power p , is in this Section resampled to 3 hour values, as defined in Section 3.2.

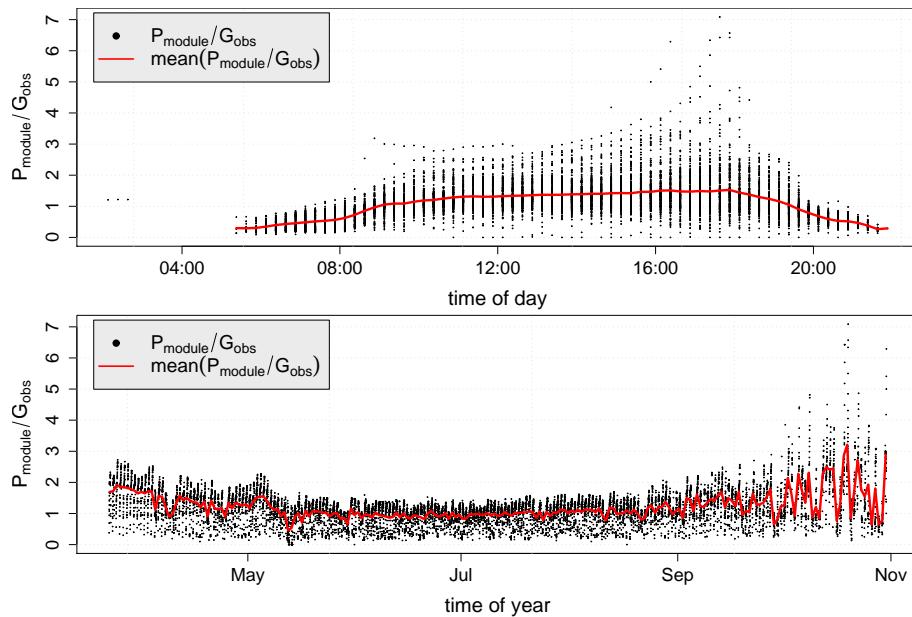


Figure 7.2: Upper plot: The ratio $\frac{P_{\text{module}}}{G_{\text{obs}}}$ versus t_{tod} . A clear indication of a changing power(G) w.r.t time of day. Lower plot: The ratio $\frac{P_{\text{module}}}{G_{\text{obs}}}$ versus time of year. Also here a clear indication of dependency is found. It is seen by a change in the distribution of the values toward spring and autumn.

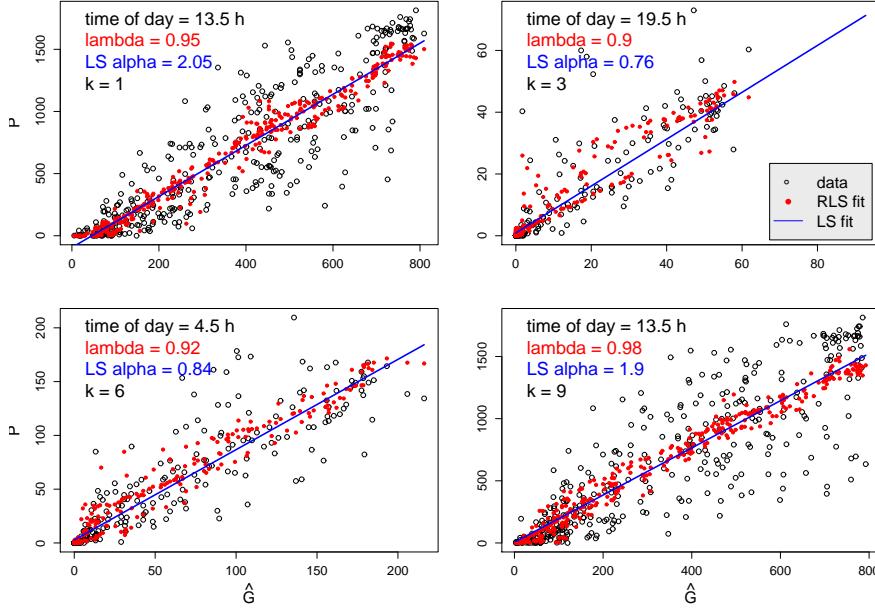


Figure 7.3: RLS fits of the power(G) for $\hat{g}_{12,k}$ where $k = 1, 3, 6, 9$. A LS fit is also shown and it is observed that α is varying (most likely as a function of time of day, but no further conclusions is made about this).

An estimation of $power(G)$ for each horizon of NWP updates at 00:00 and for each horizon of NWP updates at 12:00, is done in order to include the *time of day* dependency effect found above. The time series

$$\hat{g}_{00,k} = \{\hat{g}_{i,k}, i = 1, 3, \dots, 727\},$$

with 24 hour values holds the NWPs made at 00:00 of horizon k . The time series of p at the same time values are

$$p_{00,k} = \{p_t, t = (0 \cdot 8 + k), (1 \cdot 8 + k), \dots, (363 \cdot 8 + k)\}.$$

The linear $power(G)$ model for horizon k of NWP updates at 00:00, is then

$$\hat{p}_{00,k,t} = \beta_t + \alpha_t \hat{g}_{00,k,t}. \quad (7.1)$$

An equivalent model is made for the NWP updates at 12:00. The dependency of time of year is picked up by fitting the model with the k -step RLS with forgetting defined in Section 5.4. To get a correct result with the RLS algorithm, then p has to be lagged, for

$k \leq 8$: p is lagged 1 step (i.e. 24 hours) and 1-step RLS is used. After fitting the result $\hat{p}_{00,k}$ is lagged -1 step.

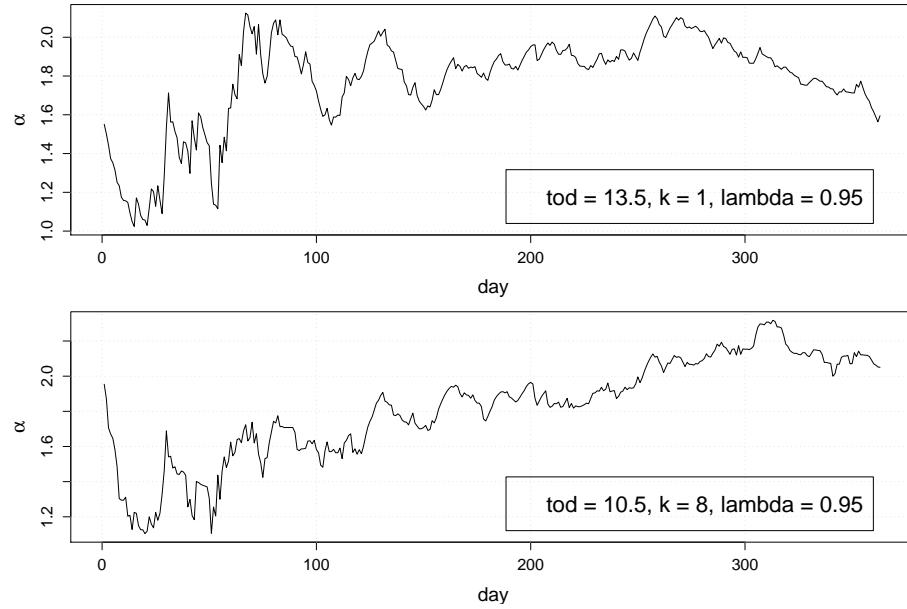


Figure 7.4: α in two RLS fits of $\hat{p}_{00,k,t} = \alpha_t \hat{g}_{00,k,t} + e_t$. A trend over the year is observed, but since only one year is used no conclusion can be made about periodicity.

$8 < k \leq 12$: p is lagged 2 steps and 2-step RLS is used. After fitting $\hat{p}_{00,k}$ is lagged -2 step.

Each RLS estimation is optimized by choosing $\lambda = 0.9, 0.91, \dots, 1$ that minimizes the *RMSE*. Results for four of the horizons are shown in Figure 7.3 and all other fits have been checked. The Least Squares (LS) fit is also shown in each plot and it is seen that α from the LS fit (the value is written on the plot) is changing considerably for different values of *time of day*. It seems that α is lower towards nighttime, but this observation needs further investigation. The fits are different for changing *time of day*, and thus a fit for each horizon is necessary. It is also noted that the λ value is lower for the two fits close to nighttime, this indicates a more varying *power(G)* over the year for *time of day* toward dawn and dusk.

The change in *power(G)* over the year is found by plotting α_t in

$$\hat{p}_{00,k,t} = \alpha_t \hat{g}_{00,k,t} .$$

as estimated by RLS over the year. It is the same model as 7.1 without the intercept. The plots for two horizons are shown in Figure 7.4. To have a realistic starting value of α the year is repeated and the plot show the second year, but of course a real conclusion about the time of year dependency can

only be made with data from more than one year. The conclusion drawn is that there is an indication of a trend in α_t over the year and thus it is found that $\text{power}(G)$ of the NWPs are changing over time.

The Auto Regressive (AR) prediction models used in this analysis assumes stationary of the modeled processes, hence time series of the τ -process is used (described in Chapter 4). In order to use the AR models and to make effective interpolation when resampling to hour values, the predicted p values in $\hat{p}_{00,k,t}$ and $\hat{p}_{12,k,t}$ is transformed to predictions of τ .

The values in $\hat{p}_{00,k}$ and $\hat{p}_{12,k}$ are transformed to similar time series with predictions of τ . This is carried through for $\hat{p}_{00,k}$ by making a time series

$$\hat{p}_{\text{cs},00,k} = \{p_{\text{cs},t}, t = (0 \cdot 8 + k), (1 \cdot 8 + k), \dots, (363 \cdot 8 + k)\},$$

of \hat{p}_{cs} resampled to 3 hour values. Now τ for horizon k can be found by

$$\hat{\tau}_{00,k} = \left\{ \frac{\hat{p}_{00,k,t}}{\hat{p}_{\text{cs},00,k,t}}, t = 1, \dots, 364 \right\}.$$

The similar transformation is made of $\hat{p}_{12,k}$ into $\hat{\tau}_{12,k}$. The transformed NWPs is then put back in time series

$$\begin{aligned}\hat{\tau}_i &= \left\{ \hat{\tau}_{00, \frac{i+1}{2}, k}; k = 1, \dots, 16; i = 1, 3, \dots, 727 \right\}. \\ \hat{\tau}_i &= \left\{ \hat{\tau}_{12, \frac{i}{2}, k}; k = 1, \dots, 16; i = 2, 4, \dots, 728 \right\}.\end{aligned}$$

Thus $\hat{\tau}_i$ is a time series with 3 hour values of predictions of τ at the same time values and with the same dimensions as \hat{g}_i .

The final steps before the NWPs are ready to be used in the prediction models, is to resample $\hat{\tau}_i$ into hour values by linear inter- and extrapolation (i.e. $\hat{\tau}_i$ is from now with hour values) and then one more step is applied to get the final time series used as input to the prediction model. Different NWPs are used for different k horizons to index $\hat{\tau}_{i,k}$ two discrete index functions are needed

$$\begin{aligned}k(t, k) &= ((t - 1) \bmod 12) + 1 + \left(\left\lfloor \frac{k - 1}{12} \right\rfloor + 1 \right) \\ i(t) &= \left\lfloor \frac{t - 1}{12} \right\rfloor + 1\end{aligned}$$

which for $k = 1, \dots, 12$ give the same result, and likewise for $k = 13, \dots, 24$ and for $k = 25, \dots, 36$. Then the NWPs used for horizon k are found by

$$\hat{\tau}_{\text{nwp},k,t} = \hat{\tau}_{i(t),k(t,k)}$$

and they are put into the time series with hour values

$$\hat{\tau}_{\text{nwp},k} = \{\hat{\tau}_{\text{nwp},k,t}, t = 1, \dots, 8760\}$$

which is used in the prediction models. As described this gives three distinct $\hat{\tau}_{\text{nwp},k}$ and they consist of the most recent NWP updates of horizons in the three intervals. To ease the notation of prediction models the k subscript is dropped

$$\hat{\tau}_{\text{nwp}} = \hat{\tau}_{\text{nwp},k} \quad (7.2)$$

and it is implicit that $\hat{\tau}_{\text{nwp},k}$ is used for horizon k .

7.2 Identification of the best ARX model

The best model with NWPs as input is identified in this section. First the model with only the NWP as input is fitted by RLS and then AR-terms are added stepwise until the performance is not clearly increased.

The model only using the NWPs as input is

$$\hat{\tau}_{t+k|t} = m + c_1 \hat{\tau}_{\text{nwp},t+k}$$

and it is named ARX_{nwp} . By doing the same optimization of λ as described in Section 6.1 on page 58, $\lambda = 1$ is chosen for the RLS fitting. The reason is most likely that the $power(G)$ was fitted with forgetting and thus the change over the year is reflected in $\hat{\tau}_{\text{nwp}}$. The first couple of horizons does show that $\lambda < 0.9$ should be used, but this is ignored.

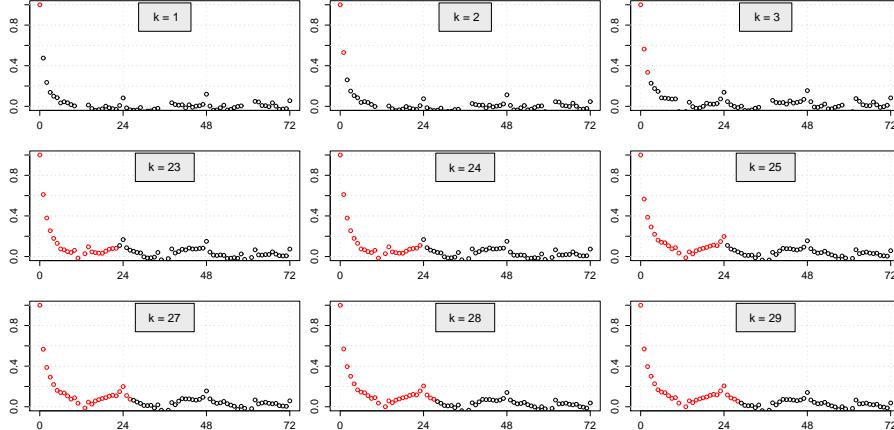


Figure 7.5: *ACF of residuals from ARX_{nwp} RLS fit of selected horizons. $k = 1, 2, 3$ Shows significance for an AR-term and peaks at lag 24,48,72 indicates a seasonal AR-term. $k = 23, \dots, 29$ show indications of a seasonal AR-term.*

The ACF of the residuals is shown in Figure 7.5 and show the need for an AR-term for the first couple of horizons. Hence this term is added and the second model fitted is

$$\hat{\tau}_{t+k|t} = m + a_1 \tau_t + c_1 \hat{\tau}_{\text{nwp},t+k}$$

and it is named $ARX_{k,k_{\text{nwp}}}$ following the naming convention described at page 58. $\lambda = 1$ optimizes nearly all horizons and is used. The ACF of the model for selected horizons is showed in Figure 7.6 and it is seen that the significance of the first couple of lags is effectively removed. The RMSE for $ARX_{k,k_{\text{nwp}}}$ is compared to ARX_{nwp} in Figure 7.7. A high positive Improvement is found in the first couple of horizons and it is decided to keep $ARX_{k,k_{\text{nwp}}}$. The ACF indicates that the seasonal lag 24 could still give an improvement to the model, so it is added

$$\hat{\tau}_{t+k|t} = m + a_1 \tau_t + b_1 \tau_{(t-24+k \bmod 24)} + c_1 \hat{\tau}_{\text{nwp},t+k}$$

this model is named $ARX_{k,k_{24},NWP}$. $\lambda = 1$ is used.

The result of the RMSE Improvement is shown Figure in 7.8 together with the model

$$\hat{\tau}_{t+k|t} = m + a_1 \tau_t + a_2 \tau_{t-1} + c_1 \hat{\tau}_{\text{nwp},t+k},$$

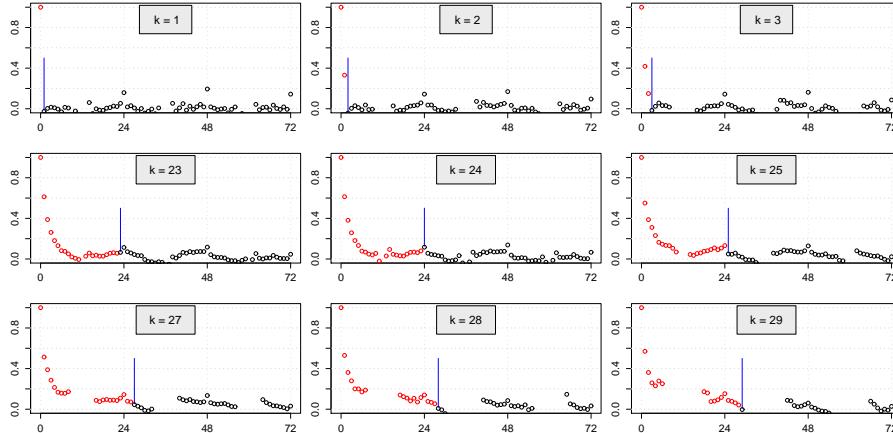


Figure 7.6: ACF of residuals from $ARX_{k,k_{\text{nwp}}}$ RLS fit of selected horizons. $k = 1, 2, 3$ show that the AR-term has removed the significance of the first couple of lags. $k = 23, \dots, 29$ still show indications of a seasonal AR-term, but it is noted that the ACF for $k = 24$ - that contains the seasonal lag - show no change for the seasonal lags compared to the ACF at $k = 24$ for ARX_{nwp} .

and it is found that neither of the added terms improve the performance enough to use them. $ARX_{k,k_{24},NWP}$ does show an mean RMSE Improvement of about 0.5%, but since there is no clear pattern and also negative RMSE Improvement for some horizons. Hence it is found that the inclusion of the 24 lag seasonal term is to overfit the data and more thorough analysis is needed to determine this clearly.

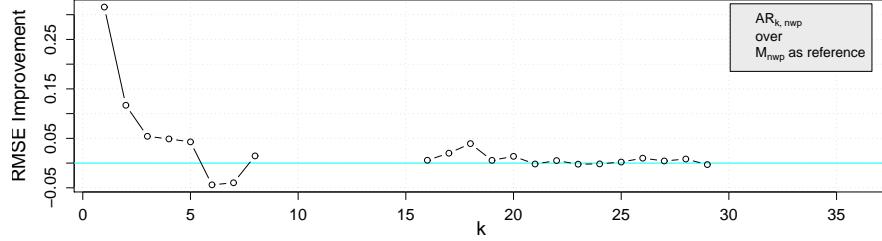


Figure 7.7: RMSE Improvement of $ARX_{k,k_{nwp}}$ over ARX_{nwp} . A clear positive Improvement of the first couple of horizons.

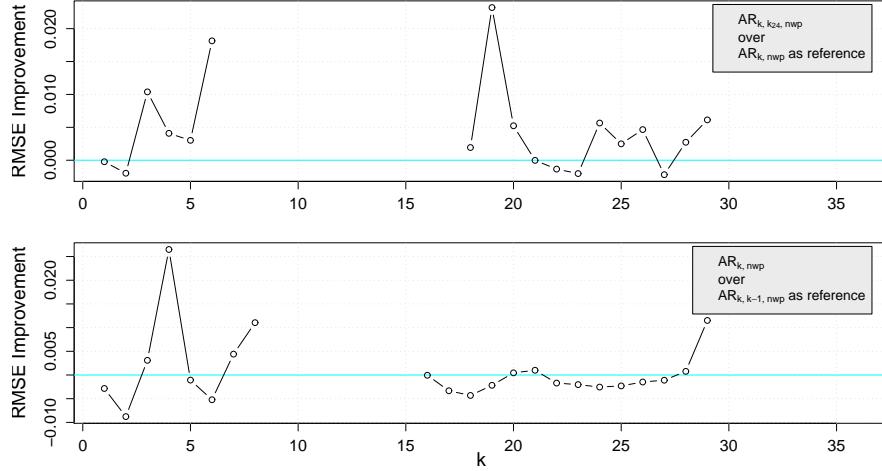


Figure 7.8: RMSE Improvement of $ARX_{k,k_{24},NWP}$ over $ARX_{k,k_{nwp}}$ and $ARX_{k,k-1,NWP}$ over $ARX_{k,k_{nwp}}$. The models are not clearly improved for all horizons, though the mean of the improvement over the horizons is positive, the terms are not added to the model.

7.3 Evaluation of the selected ARX model

The selected model using NWPs as input is

$$\hat{\tau}_{t+k|t} = m + a_1 \tau_t + c_1 \hat{\tau}_{\text{nwp},t+k}$$

denoted by $ARX_{k,k_{\text{nwp}}}$, is evaluated in this section. All result are concerning τ and when the Reference model is used it is transformed by \hat{p}_{cs} .

The plot in figure 7.9 show RMSE, BIAS and Completeness for all horizons. The forecasts are clearly improved compared to RMSE of the Reference model. The poor performance of $k = 18$ are attributed to the weak performance of the Clear Sky model toward dawn and dusk. This also holds for the BIAS.

The cumulated prediction errors are plotted for six selected horizon in Figure 7.10. Some remarkable changes in short periods are observed, and this should lead to further investigations of what causes this effect. Maybe some repeated phenomenon of all periods can be found, and variables containing this information could be included in the model. It should also be investigated if some information to be used for forecasts accuracy, can be found. This figure is also interesting to compare with the similar figure for the selected $AR_{k,k_{24}}$ model without NWP input.

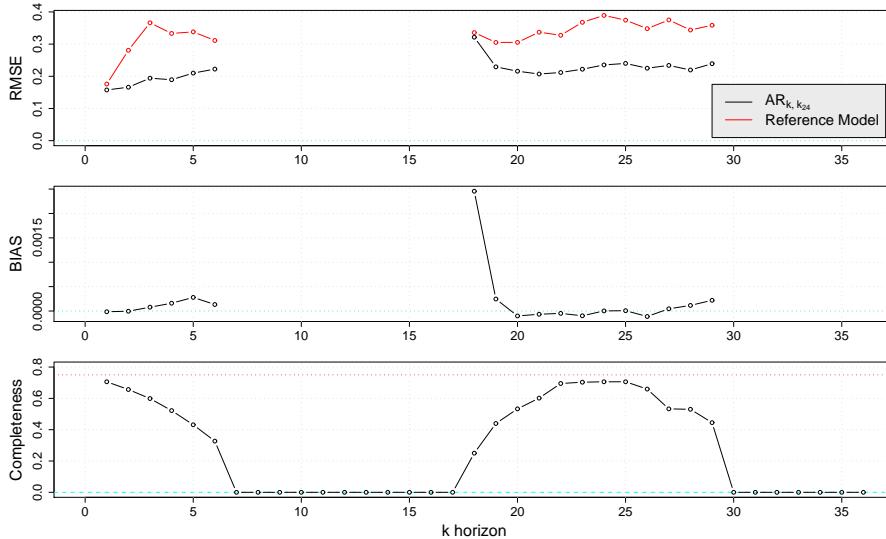


Figure 7.9: Upper: The RMSE of $ARX_{k,k_{\text{nwp}}}$ and the Reference model. Middle: BIAS of $ARX_{k,k_{\text{nwp}}}$ for each horizon. Lower: The Completeness. Note that the Completeness only scale up to 0.75. Due to the length of the burn-in period the first 25% of the predictions are removed.

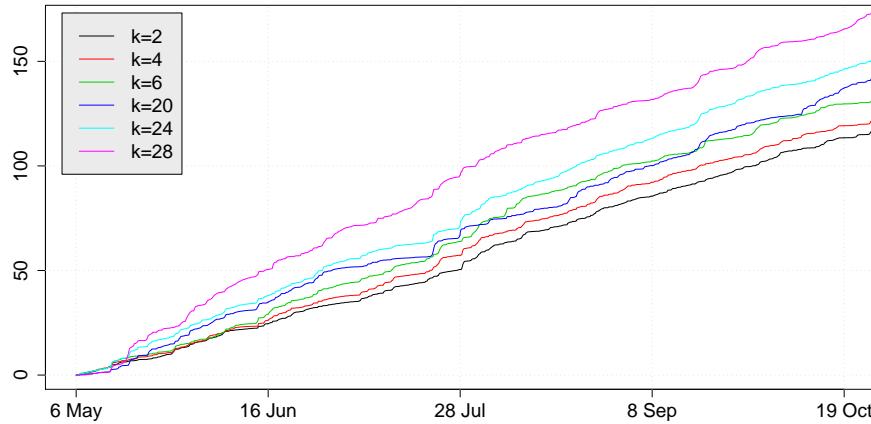


Figure 7.10: *The cumulated prediction errors of $ARX_{k,k_{nwp}}$. The profound changes in short periods should definitely lead to close exploration of the causes.*

A quick visual inspection is done and it is found that the short periods seems to be the same, but it is noted that differences are more profound for the $ARX_{k,k_{nwp}}$ model.

The histograms in Figure 7.11 show estimates of the distribution of the errors for four selected horizons. The errors are well centered around 0 and the sharpness is decreasing as the horizon is increasing. QQplots have also been checked for each horizon and no indications of weak points in the model is found.

Figure 7.12 is a plot of the prediction errors as a function of the predicted value $\hat{\tau}$, and the BIAS for 10 equal sized bins over the range of $\hat{\tau}$ is shown by a red line. The scale of the BIAS is in the right side of the plot. The plot gives a view of how the distribution of the errors are conditioned on $\hat{\tau}$. The errors is seen to be well centered around 0 and the errors are within the same kind of band as described on in Section 6.1 on page 64. The BIAS of the different bins does not indicate any clear patterns, though it is found that for $k = 4$ and $k = 18$ the levels of BIAS are highest. This is also reflected in Figure 7.9, where horizons toward dawn and dusk have a higher BIAS. This effect is attributed to the decreasing performance of the Clear Sky model toward dawn and dusk. As for the selected AR model, it is also seen quite clearly that the deviation of the errors are decreasing toward the limits of $\hat{\tau}$, and this information must be used when modeling the uncertainties of the forecasts.

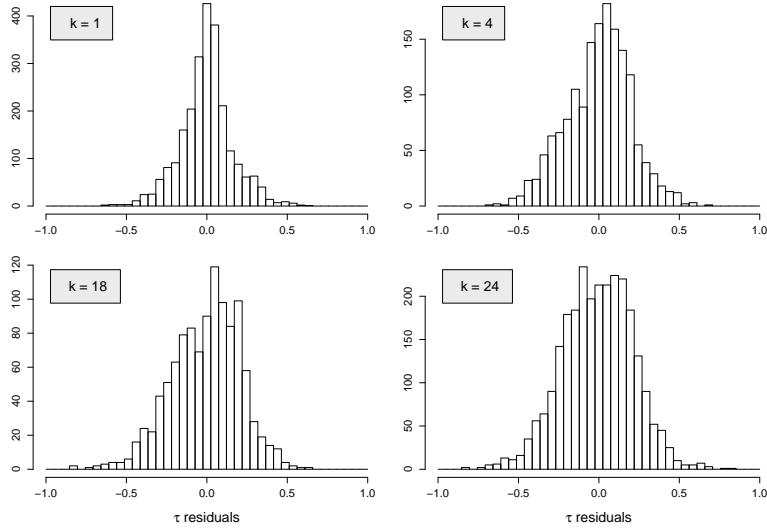


Figure 7.11: The error distribution of four horizons for $ARX_{k,k_{\text{nwp}}}$. No indications of severe skewness is seen, though $k = 4$ and $k = 18$ are leaning a bit toward one side.

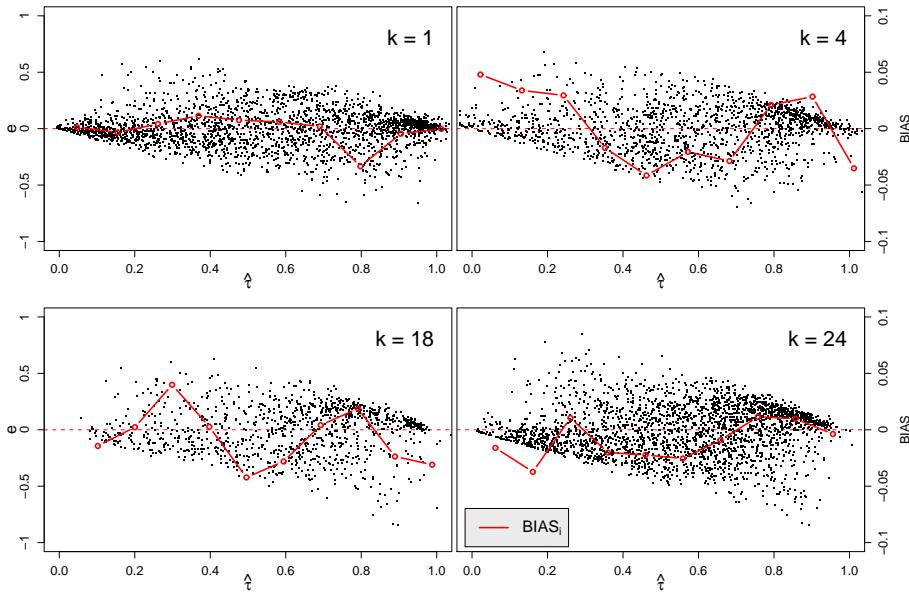


Figure 7.12: The distribution of the prediction errors conditioned on $\hat{\tau}$. The red line shows the $BIAS_i$ for 10 bins uniformly dividing the range of $\hat{\tau}$.

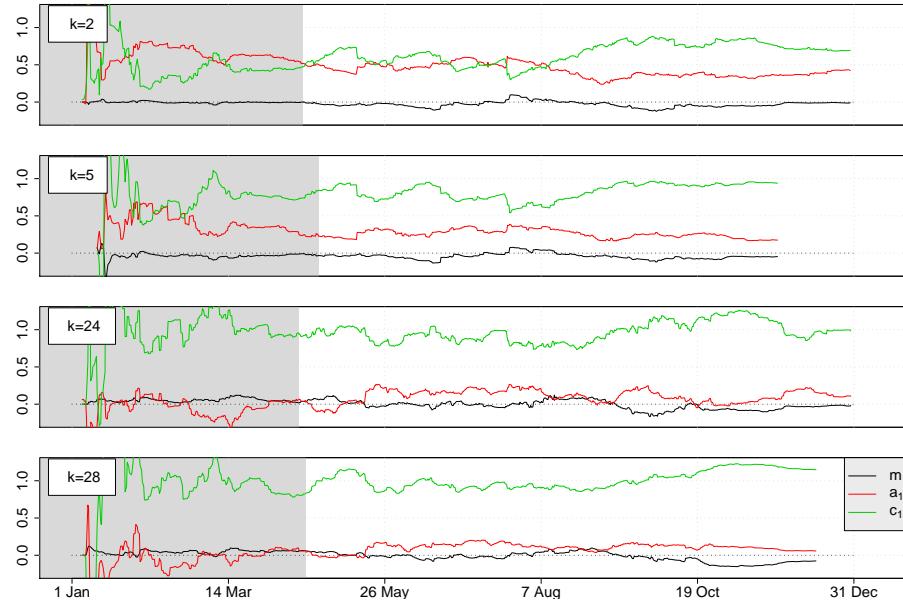


Figure 7.13: The coefficient estimates of $ARX_{k,k_{\text{nwp}}}$ seen over time for four selected horizons. The burn-in period is marked with a grayed background. $\lambda = 0.995$ is used. It is seen that for the two long horizons only the information in the NWP input term is utilized.

Finally the plot in Figure 7.13 shows the estimated values of the coefficients in $ARX_{k,k_{\text{nwp}}}$ over time. The chosen optimal level of $\lambda = 1$ (i.e. no forgetting) result in smooth curves with little variation, thus in order to allow for some variation and to be able to compare with the similar plot for $AR_{k,k_{24}}$, the same level $\lambda = 0.995$ is used in both plots. It is seen that there is less variation of the curves compared to the similar plot for the selected AR model. Concerning the levels of the coefficients there is a change over the horizons. For the shortest horizons the information in the AR-term is utilized, whereas for the longer horizons only the NWP input is used. This change is reflected from the RMSE performance plot in Figure 7.9. It is also seen that the intercept coefficient is close to 0 and could be left out of the model.

CHAPTER 8

Summary and discussion

In the first section the final evaluation with solar power predictions is carried out for the two selected model. Then follows the calculation of performance summary numbers for very short-term horizons and for next days horizons. Then a simple approach to uncertainty modeling of the forecasts are outlined, and this is applied when examples of solar power forecasts are shown. The chapter is ended with a discussion of the developed forecasting method and ideas for further work and refinements.

8.1 Final solar power forecast evaluation

The performance of the two selected models are evaluated and compared to the Reference model. The evaluation is done with the solar power forecast matrix $\hat{\mathbf{P}}$ defined in Section 5.1 and the error measures defined in Section 5.2.1. The $NRMSE_k$ is calculated for each horizon. Note that the $Completeness_k$ is the same for predictions of p as for the predictions of τ , and this has been evaluated for both the selected models. The result for $AR_{k,k_{24}}$ is seen in Figure 6.6 and for $ARX_{k,k_{\text{nwp}}}$ in Figure 7.9.

The plot in Figure 8.1 shows the results for $AR_{k,k_{24}}$. It is found that a distinct improvement over the Reference model achieved. The $I_{NRMSE,k}$ for $k = 1, 2$ is low because the Reference model has a good performance for the very short horizons. It is noted that much higher $I_{NRMSE,k}$ for very short horizons are found

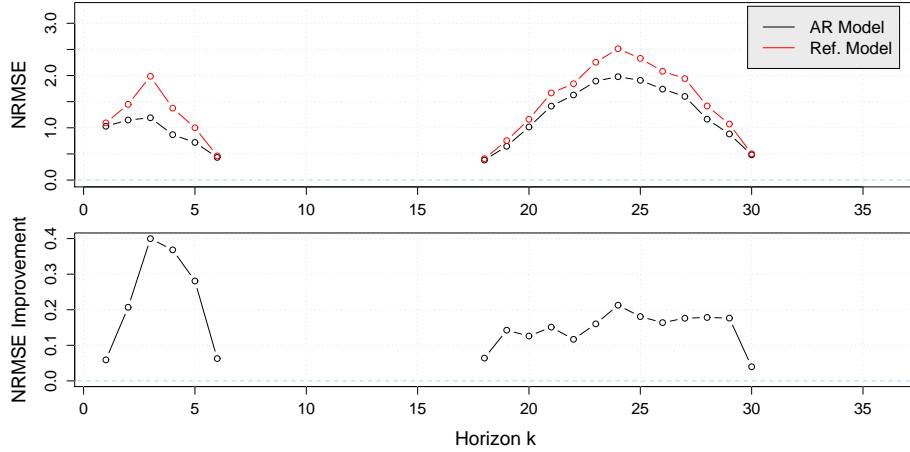


Figure 8.1: *Evaluation of $AR_{k,k_{24}}$.* The upper plot show the $NRMSE_k$ and the lower plot the $I_{NRMSE,k}$ over the Reference model. A distinct improvement over the Reference model is seen.

in other studies using similar persistence model as reference. The study [21] finds up to $I_{R\text{MSE},k=1} = 74\%$ using hour values. This is because the predictions are made in a sliding window, i.e. at all steps throughout the day, whereas in the present study only predicted values close to noon are included. Since the mean level of the solar power changes throughout the day, and since this change is smaller around noon than the overall level of change, a one step persistence model will perform much better around noon. Hence the results for $k = 1$ in this study and in [21] can not be compared. Generally it is noted, that there can be a huge difference for a given model in performance between forecasts only produced at one *time of day*, and forecasts produced in a sliding window.

The evaluation result for $ARX_{k,k_{\text{nwp}}}$ is shown in Figure 8.2. A clear improvement is observed for all horizons, though the same effect for $k = 1, 2$ as described above is seen.

A comparison between $ARX_{k,k_{\text{NWP}}}$ and $AR_{k,k_{24}}$ is done and the result is shown in Figure 8.3. It is seen that $ARX_{k,k_{\text{NWP}}}$ has a better performance for all horizons, but it is especially superior for the next day horizons $k = 18, \dots, 29$. Considering the small $I_{NRMSE,k}$ for next day horizons contributed by the k AR-term, seen in Figure 7.7, and the fact that the k AR-term is in both models compared here. It is found that a model only using only NWPs as input would give the same $I_{NRMSE,k}$ as seen in Figure 8.2, in the magnitude of 20% to 30%. For shorter horizons the results indicate that the k AR-term is needed to improve the performance especially at horizons less than two hours, but that NWPs does give a considerable improvement for horizons of 2 hours and above.

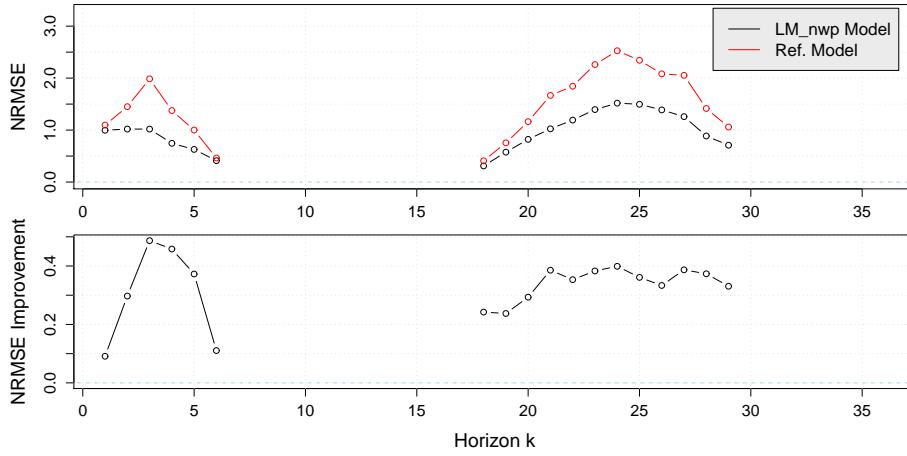


Figure 8.2: *Evaluation of $ARX_{k,k_{\text{NWP}}}$.* The upper plot show the $NRMSE_k$ and the lower plot the $I_{NRMSE,k}$ over the Reference model. A positive $I_{NRMSE,k}$ over the Reference model is seen, up to 50 % for very short-term and around 35 % for next day horizons.

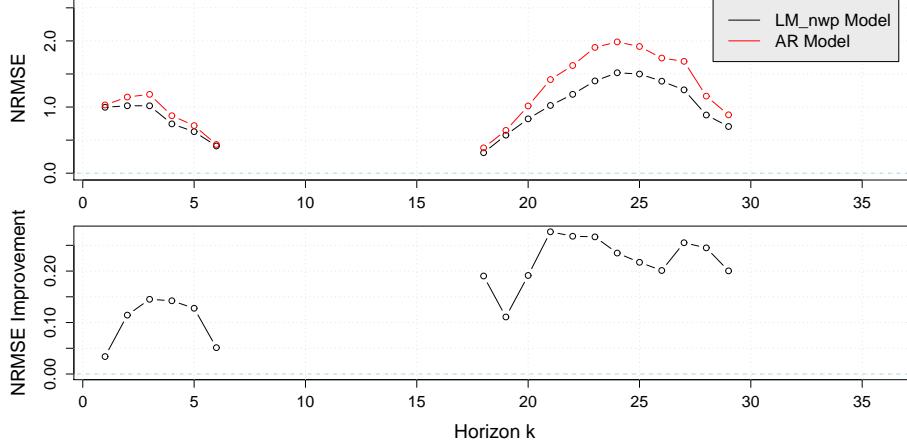


Figure 8.3: *A comparison of $ARX_{k,k_{\text{NWP}}}$ to $AR_{k,k_{24}}$ as a reference.* It is seen that using NWPs as input the performance is improved for all horizons, but especially for the next day horizons $k = 18, \dots, 29$ a considerable improvement is achieved.

8.2 Performance summary numbers

Summary numbers of the performance is calculated. The solar power forecast matrix $\hat{\mathbf{P}}$ defined in Section 5.1 and the error measures defined in Section 5.2.1 are used. The summary numbers are calculated separately for very short-term horizons $k = 1, \dots, 6$ and for next day horizons $k = 19, \dots, 29$. All available predicted values of the given horizons are included. The summary numbers are calculated for both the two selected models. The summary numbers are based on $NRMSE$ and I_{NRMSE} over the Reference model. It is noted that the error measures are only calculated for values predicted by both selected models, and therefore the $Completeness_k$ of the very short-term and next day horizons, are also relevant to evaluate. It is calculated as

$$Completeness_{k=k_s:k_e} = \frac{1}{0.75} \frac{1}{363(k_e - k_s + 1)} \sum_{k=k_s}^{k_e} \text{finite}(\hat{\mathbf{P}}_{1:363,k})$$

where $\text{finite}(\hat{\mathbf{P}}_{1:363,k})$ is the number of finite (i.e. predicted) values at horizon k . The division by 0.75 is done because of the burn-in period in RLS, where 25% of the predicted values are removed. $k_s = 1$ and $k_e = 6$ for the very short-term horizons, and $k_s = 19$ and $k_e = 29$ for the next day horizons. The results are

$$\begin{aligned} Completeness_{k=1:6} &= 72\% \\ Completeness_{k=19:29} &= 79\%. \end{aligned}$$

The $NRMSE_k$ of the $AR_{k,k_{24}}$ model are

$$\begin{aligned} NRMSE_{k=1:6} &= 0.99 \\ NRMSE_{k=19:29} &= 1.58 \end{aligned}$$

and for the $ARX_{k,k_{NWP}}$ model

$$\begin{aligned} NRMSE_{k=1:6} &= 0.88 \\ NRMSE_{k=19:29} &= 1.20. \end{aligned}$$

The improvement over the Reference model is for the $AR_{k,k_{24}}$ model

$$\begin{aligned} I_{NRMSE,k=1:6} &= 28\% \\ I_{NRMSE,k=19:29} &= 17\% \end{aligned}$$

and for the $ARX_{k,k_{NWP}}$ model

$$\begin{aligned} I_{NRMSE,k=1:6} &= 36\% \\ I_{NRMSE,k=19:29} &= 37\%. \end{aligned}$$

As discussed in Section 8.1, the improvement achieved using NWPs as input to the prediction models, is found by the improvement of $ARX_{k,k_{\text{NWP}}}$ compared to $AR_{k,k_{24}}$ as a reference

$$\begin{aligned} I_{\text{NRMSE},k=1:6} &= 11\% \\ I_{\text{NRMSE},k=19:29} &= 24\%. \end{aligned}$$

Finally it is noted, that the I_{NRMSE} of the Reference model including all values over the Reference model only including the values also predicted by the two selected models (i.e. the values used in the calculations above), is

$$\begin{aligned} I_{\text{NRMSE},k=1:6} &= 36\% \\ I_{\text{NRMSE},k=19:29} &= 7\%. \end{aligned}$$

Thus it is seen that including all values results in a lower $RMSE$. This result only show that the values removed (they are mainly toward dawn and dusk) do not change the performance negatively. It is thus found that these values would not change the improvement results of the two models over the Reference model very much.

8.3 Uncertainty modeling

Modeling the uncertainties of the solar power forecasts increases their usefulness considerably. Since a thorough study of the uncertainties is beyond the scope of the present thesis, just a simple approach is outlined. Quantile regression is used inspired by [17]. In Figure 8.4 a plot of $\tau_{\text{mat}}[:, k]$ versus $\hat{\tau}_{\text{mat}}[:, k]$ is shown for horizons $k = 1$ and $k = 24$. The red lines are estimated $\hat{Q}(\tau_q, \hat{\tau})$, that are quantiles of the probability distribution function at a given value of $\hat{\tau}$.

A 1D version of the smoothing kernel in Section 4.2.1 is used. The quantile $\hat{Q}(\tau_q, \hat{\tau})$ is estimated for $\tau_q = 0.05, 0.25, 0.50, 0.75$ and 0.95 . It is seen that the uncertainties are lower for $\hat{\tau}$ toward 0 and 1, than in the middle around 0.5.

The quantiles for each power prediction value $\hat{p}_{i,k}$ can thus be found by

$$\hat{p}_{\tau_q, i, k} = \hat{Q}(\tau_q, \hat{\tau}_{i, k}) \cdot \hat{p}_{\text{cs}, t+k}.$$

These are calculated with $\tau_q = 0.05$ and $\tau_q = 0.95$ for each forecast.

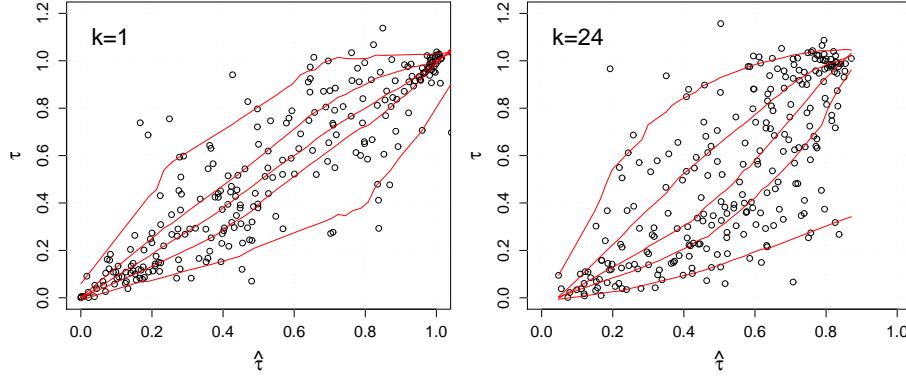


Figure 8.4: Quantile estimates $\hat{Q}(\tau_q, \hat{\tau})$ for $\tau_q = 0.05, 0.25, 0.50, 0.75, 0.95$ and horizons $k = 1$ and $k = 24$. A bit sparse data gives a bit wickedly quantile estimates, but since only a simple approach is used this is accepted.

8.4 Solar power forecasts for Elspot

The final result are solar power forecasts with uncertainty estimates. The forecasts are as explained, issued at 12:00 UTC every day and consists of hour value predictions. Forecasts issued at the day of year $t_{\text{day}} = 120, 165, 210, 255, 300$ are shown in Figure 8.5 for $AR_{k, k_{24}}$ and in Figure 8.6 for $ARX_{k, k_{\text{nwp}}}$. The two models give almost the same forecasts for horizon $k = 1, \dots, 6$, but $ARX_{k, k_{\text{nwp}}}$ using NWPs as input, are clearly better for next day horizons.

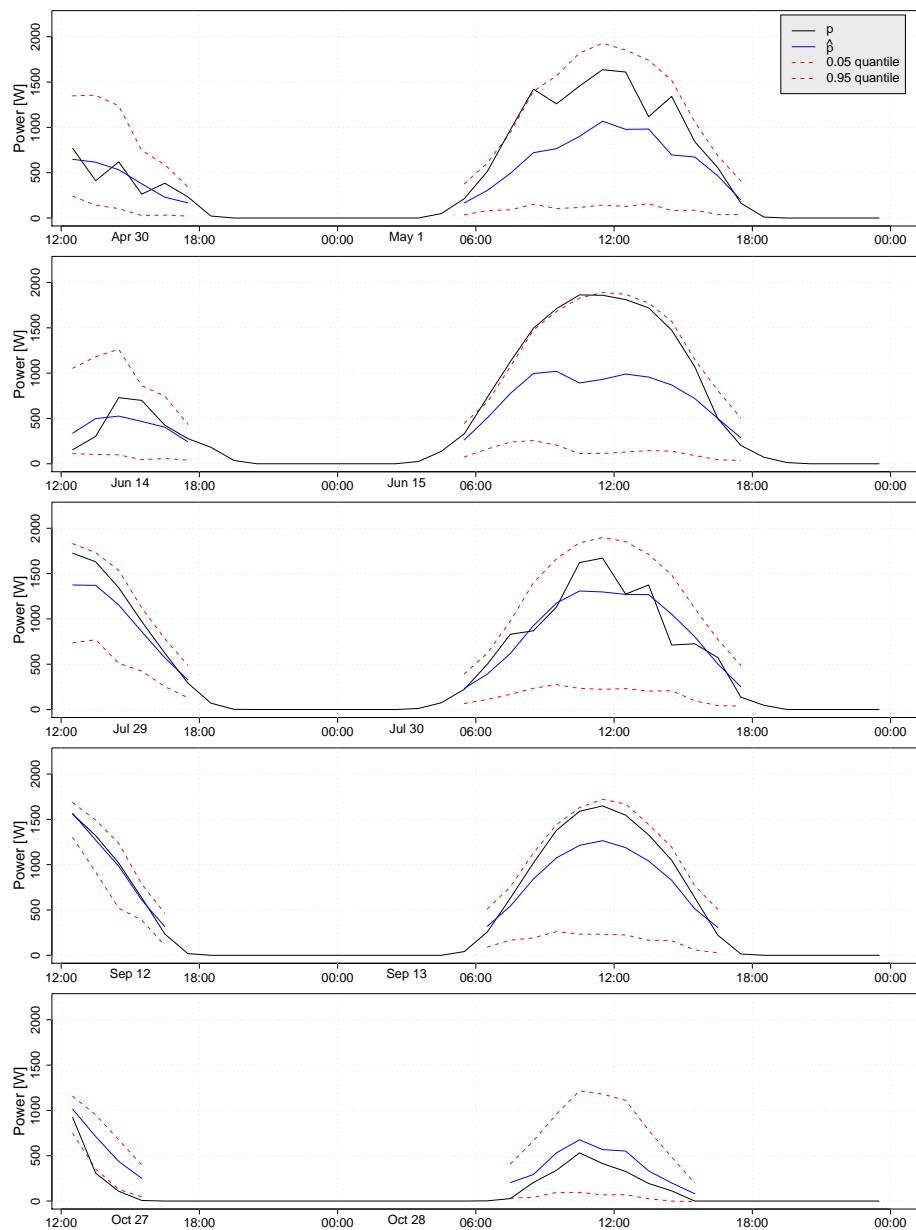


Figure 8.5: *Forecasts made with $AR_{k,k_{24}}$ for $t_{\text{day}} = 120, 165, 210, 255, 300$.*

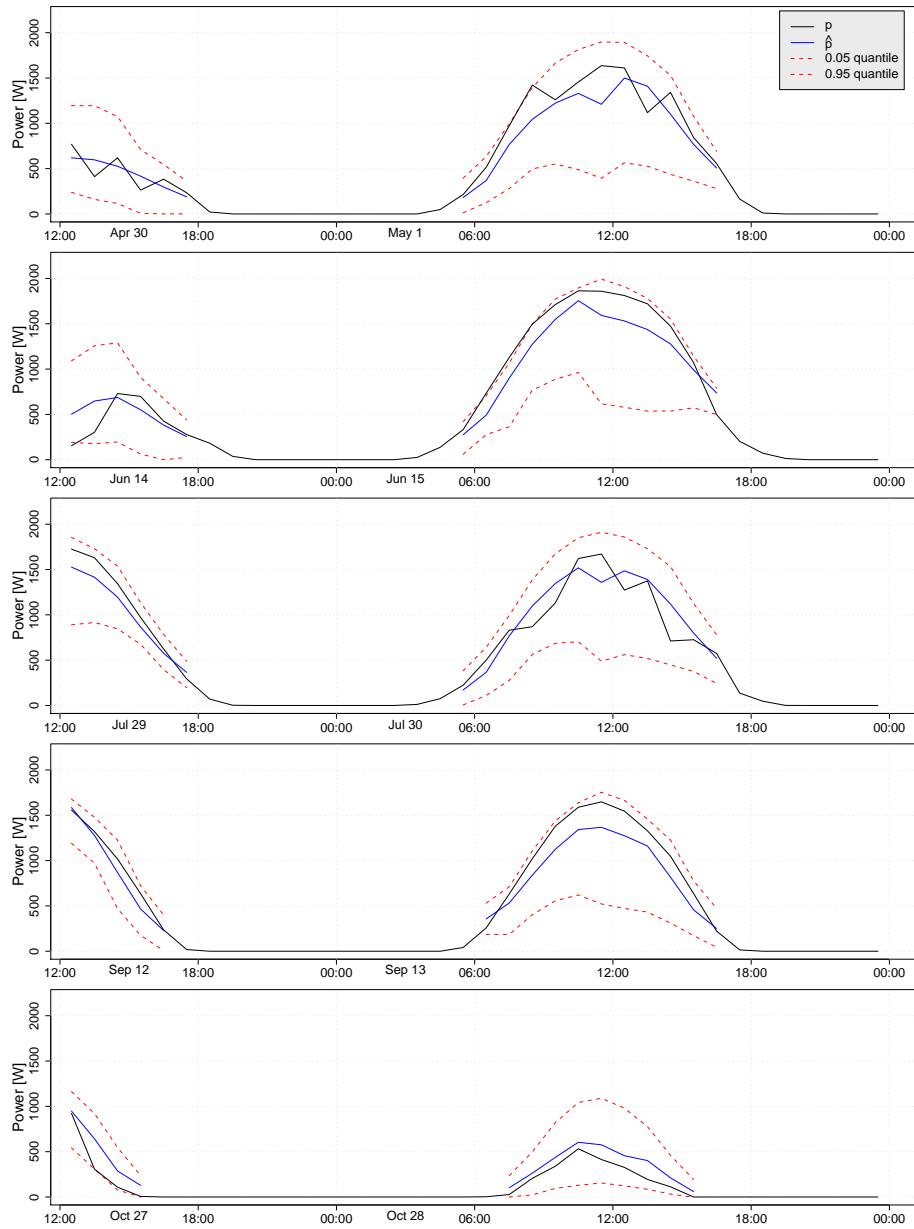


Figure 8.6: Forecasts made with $ARX_{k,k_{\text{nwp}}}$ for $t_{\text{day}} = 120, 165, 210, 255, 300$.

8.5 Discussion and further work

In this section the findings throughout the modeling process are discussed and ideas for further work are outlined.

8.5.1 The present method

The present method for solar power forecasting has - inspired by previous studies - been developed from scratch. A new approach to make a clear sky model has been proposed, and an adaptive prediction model based on RLS make up a solid framework allowing for model extensions and further refinements.

8.5.2 Results and model evaluation

The results achieved by the present method are reasonably good, but most likely a bit optimistic, since a strict cross-validation have not been conducted. It is noted though, that RLS only utilize past values and thus do not use any degrees of freedom. Only the Clear Sky model and the optimization of λ use degrees of freedom. To which extend this is reflected in the results, should be investigated in further work by doing cross-validation. Similarly the models should be evaluated on data from other PV systems, e.g. to investigate how local weather phenomenons at other locations will influence the model identification and performance.

8.5.3 Model extensions with MOS

One of the most obvious extensions of the method, is to include more NWP variables in the model of $power(G)$ made in Section 7.1. The German study [9] reports that the general configuration of NWP models are not optimal w.r.t global irradiance. They report that the solar power forecasts can be improved by changes in the configuration of the NWP model, and by applying model output statistics (MOS), which is a technique where climate variables from the output of the NWP model is used as input to a model of a climate variable. The modeling of $power(G)$ resembles MOS and an extension with: cloud cover indexes, temperature, wind speed and wind direction, should be carried out. NWPs of these variables are already in the dataset given by DMI. Furthermore a deterministic translation of global irradiance to irradiance in the plane of the PV module, can be applied as in [14].

8.5.4 Further investigation of periods with high errors

Further investigation of the models in periods where a high level of errors are observed should be carried out. This can reveal weak points in the models and give ideas for refinements to the method.

8.5.5 Improved statistical techniques for the Clear Sky model

The performance of the Clear Sky model should be improved especially for the estimates toward dawn and dusk. Refinements such as varying the parameters in different parts, and the development of a method to optimize the estimations quantitatively, could be the next thing to do. The improvements of more advanced statistical techniques, such as two-dimensional local polynomials and two-dimensional splines should be explored. Data from more years should be included, and the estimation should be made adaptively and only past values should be used as input.

8.5.6 Comparison to other developed methods

Several forecasting methods have been developed in other studies, see Chapter 2. It is found that it is hardly possible to compare the performance of the present method to the others. Two main reasons complicate this task. First of all different datasets are used in the studies. These are both different w.r.t. size and resolution, but also varying weather conditions influences performance results. Secondly different error measures are used. To make comparison between methods from different studies possible, a common framework for evaluation of solar power forecasting methods should be developed. This could be inspired by a similar project done in [16] for wind power. Furthermore a project where datasets used in different studies are collected and made available for everybody, would enhance the possibilities to compare methods in a much more informative way.

8.5.7 An operational model

Work must done before an operational method is ready. The completeness should be increased by refining the Clear Sky model toward dawn and dusk and in the winter period, and it should modified so that only input of past values are used. Handling of missing input should be dealt with, so that e.g. if an input to the RLS is missing, the method should still give a prediction and adapt to the new conditions. Several other data handling issues will arise if an operational

version of the method is going to be implemented. A substantial problem already exists. In the current configuration, a GSM modem at each PV system is called once a day to acquire the solar power observations to a centralized server. This of course makes it impossible to use this data for very short horizons - where it improves the result considerably. Though less disastrous for next day horizons, it will influence the MOS etc. for these horizons. It is possible to connect the PV systems online via the Internet and acquire the data much faster, but if this can be done with the current implemented GSM modems needs to be investigated.

The opportunities of producing solar power forecasts for Elbas, can be explored with the current dataset. It will be possible to make very short-term solar power forecasts with quarter values in a sliding window. Thus a study focusing on these types of forecasts could be carried out. When forecasting less than 6 hours in a sliding window, a method by [9] analysing satellite images performs very good. Cloud indexes from images can be used as input to the present method. The images used are from the METEOSAT and they also cover Denmark. The use of these should be investigated.

In a future scenario with an increased number of grid-connected PV systems installed on rooftops evenly scattered all over Denmark, the spatial averaging effect of the forecasts of the entire ensemble, is of high interest. Results from [14] shows an overall reduction factor of 0.4-0.5 for a region of the size of Germany.

With the Clear Sky model, estimates of a quantity very alike to the cloud cover index found from satellite images can be calculated for the geographic position of the PV system. This information would reassemble the same spatial information of cloud cover as found in a satellite image. It is imagined that this data in a very good temporal resolution could be utilized together with e.g. satellite images and MOS, to give very precise solar power forecasts for very short horizons. Especially with the spatial averaging effects of the overall solar power of a PV ensemble covering a large region, it is expected that some time in the future such forecasts can be produced with very high accuracy.

8.5.8 Further work

The most obvious further refinements to the method:

- Data from more years should be used to find seasonal periodicity and to carry out cross-validation.
- Modifying the Clear Sky model, so that only historical values are used, i.e. making it adaptive by a recursive update.
- Using NWPs of other climate variables as input to the models. NWPs of cloud cover index, temperature, wind direction, and wind speed, are already in the dataset.

CHAPTER 9

Conclusion

A method that predicts solar power has been developed. It issues solar power forecasts each day at 12:00 UTC with hour values at horizons 1 to 36 hours. The total power output of 21 PV systems within an area of a few square kilometers, has been used for the analysis. The data cover 2006. The thesis is focused on forecasts for trading on the Nordic electric power market Elspot. NWPs of global irradiance from the mesoscale model DMI-Hirlam are used. They are updated at 00:00 UTC and 12:00 UTC each day and they are with 3 hour values.

A clear sky model has been developed and it is used to transform the solar power process into a stationary process τ_P . This process is alike the transmittance of the atmosphere and a realization τ of τ it is used as input to prediction models. One model is fitted for each horizon k and parameters are estimated adaptively. Finally the τ predictions are transformed back to predictions of solar power. A reference model with the best naive predictor for each horizon is found and used for model evaluation. The final evaluation of the present method is done by *NRMSE*, where the normalization is done with the global mean of the solar power including nighttime values.

Two prediction models are identified: one using only past observations of solar power and another also utilizing NWPs. The best model only using past observations is identified as

$$\hat{\tau}_{t+k|t} = m + a_1 \tau_t + b_1 \tau_{(t-24+k \bmod 24)} .$$

It achieves an total *NRMSE* of 0.99 for horizons less 6 hours and a *NRMSE* of

1.58 for 19-29 hours. This is a *NRMSE* improvement of 28 % for less than 6 hours and 17% for 12-36 hours over the Reference model.

The best prediction model also utilizing NWPs as input are identified as

$$\hat{\tau}_{t+k|t} = m + a_1 \tau_t + c_1 \hat{\tau}_{\text{nwp}, t+k} .$$

In the final evaluation it achieves a *NRMSE* of 0.88 for horizons less than 6 hours and a *NRMSE* of 1.2 for 19-29 hours. This a *NRMSE* improvement of 36% for less than 6 hours and 37% for 19-29 hours over the Reference model.

It is noted that strict cross validation have not been done. It is also noted that only a small number of degrees of freedom are used by the method, since the predictions are done with recursive least squares (RLS), where only past values for updating of parameter estimates are used. The Clear Sky model and optimization of λ used in RLS, use few degrees of freedom.

The results clearly indicate that utilizing NWPs improves the results considerably for next day horizons, but for horizons less than 6 hours online solar power observations are a requisite to optimize the quality of the forecasts. Hence it is found that a method for solar power forecasting focusing on trading at Elspot, should utilize NWPs. For such a method solar power observations should at least be acquired once a day to be used for statistical analysis.

A simple approach of modeling the uncertainties of the solar power forecasts is outlined. It is found that the highest precision of forecasted solar power is achieved when overcast or clear sky is predicted, whereas predictions of a cloud condition in between yield high uncertainties.

Several parts of the present method have to be further developed before it can be applied in operational environments. Data acquisition and preprocessing is an issue, and handling missing inputs should also be dealt with. Further development of the method so that full completeness of the forecasts is achieved, is also a requisite.

To improve the performance of the present method, many parts of it can be refined and extended. Further work of refinements to the present method is suggested along with ideas of extensions. It has been found that it is very difficult to do performance comparisons between studies of solar power forecasting methods. Ideas of a framework for standardizing the evaluation of solar power forecasting methods are therefore suggested.

Bibliography

- [1] Cano D., Monget J., Albuisson M., Guillard H., Regas N. and Wald L.: *A method for the determination of the global solar radiation from meteorological satellite data.* Solar Energy, 37, p. 31-39, (1986), Internet:
http://www.helioclim.org/publications/1982_SE_Cano.pdf.
- [2] Chowdhury B. and Rahman S.: *Forecasting sub-hourly solar irradiance for prediction of photovoltaic output.* IEEE Photovoltaic Specialists Conference, New Orleans, USA, pp. 171-176, (1987).
- [3] Crispim E.M., Ferreira P.M. and Ruano A.E.: *Solar radiation prediction using RBF Neural Networks and cloudiness indices.* International Joint Conference on Neural Networks, Vancouver, BC, Canada July 16-21, (2006).
- [4] Danish Meteorological Institute: *The DMI Weather forecasting system,* Internet:
http://www.dmi.dk/eng/index/research_and_development/dmi-Hirlam-2.htm and <http://Hirlam.org/>,
(2008-01-17).
- [5] Enfor: Forecasting and Optimization for the Energy Sector. Internet:
<http://www.enfor.eu>.
- [6] Girodo M. (2006): *Solarstrahlungsvorhersage auf der Basis numerischer Wettermodelle,* PhD thesis, Oldenburg University (2006).
- [7] Glahn H.R., Lowry D.A. (1972). *The use of model output statistics (MOS) in objective weather forecasting.* Journal of Applied Meteorology, 11, 1203-1211.

- [8] Hastie T., Tibshirani R. and Friedman J.: *The elements of statistical learning*, Springer Series in Statistics (2001).
- [9] Heinemann D., Lorenz E. and Girodo M.: *Forecasting of Solar Radiation*. In: Dunlop E.D., Wald L., Suri M. (Eds.): Solar Energy Resource Management for Electricity Generation from Local Level to Global Scale. Nova Science Publishers, Hauppauge (2006).
- [10] Hokoi S., Matsumoto M. and Kagawa M. (1990): *Stochastic models of solar radiation and outdoor temperature*. ASHRAE Trans. 2, 245-252.
- [11] Honsberg C. and Bowden S.: *Photovoltaics CDROM*, Solar Hydrogen IGERT, Internet:
<http://www.udel.edu/igert/pvcdrom>, (2008-01-17).
- [12] Kjær S.B.: *Solcelleteknologiens status i Danmark*. Institut for Energiteknik, Aalborg Universitet, Internet:
<http://www.iet.auc.dk/sec3/solceller/pdf/ELTEKNIK.pdf>.
- [13] King D.L., Boyson W.E. and Kratochvil J.A.: *Photovoltaic Array Performance Model*, Sandia National Laboratories, November 2003, Internet:
<http://www.mauisolarsoftware.com/MSESC/PerfModel2003.pdf>,
(2008-01-17).
- [14] Lorenz E., Heinemann D., Wickramarathne H., Beyer H.G. and Bofinger S.: *Forecast of Ensemble Power Production by Grid-connected PV Systems*, Proc. 20th European PV Conference, September 3-7, 2007, Milano (2007).
- [15] Madsen H.: *Time Series Analysis*, Chapman & Hall/CRC, (2007).
- [16] Madsen H., Kariniotakis G., Nielsen H.Aa., Nielsen T.S. and Pinson P.: *A Protocol For Standardizing The Performance Evaluation Of Short-Term Wind Power Prediction Models*, Wind Engineering Volume 29, no. 6, (2006).
- [17] Møller J.K.: *Modeling of Uncertainty in Wind Energy Forecast*, IMM-2006-10, Technical University of Denmark, Internet:
http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/4428/pdf/imm4428.pdf,
(2008-01-17).
- [18] Nielsen T.S., Madsen H., Nielsen H.Aa., Pinson P., Kariniotakis G., Siebert N., Marti I., Lange M., Focken U., Bremen L. von, Louka P., Kallos G. and Galanis G.: *Short-term Wind Power Forecasting Using Advanced Statistical Methods*, ANEMOS, (2006), Internet:
http://anemos.cma.fr/download/publications/pub_2006_paper_EWEC06_WP3statistical.pdf,
(2008-02-02).

- [19] Nord Pool. The nordic power market, Internet:
<http://www.nordpool.com>
<http://www.nordpoolspot.com/PowerMaket/>.
- [20] PVresources.com: *Photovoltaic symbols and glossary*, Internet:
<http://www.pvresources.com/en/symbols.php>
<http://www.pvresources.com/en/glossary.php>, (2008-02-02).
- [21] Sfetsos A. and Coonick A.H.: *Univariate and Multivariate Forecasting of Hourly Solar Radiation with Artificial Intelligence Techniques*. Solar Energy Vol. 68, No. 2, pp. 169-178, (2000).
- [22] Energimidt: *Sol300 and sol1000*, Internet:
<http://www.sol300.dk>
<http://www.energimidt.dk/emweb/Privat/>
Alternativ+energi/Solceller/, (2008-1-17).
- [23] PV Resources: *List of largest PV power plants and further PV information*, Internet:
<http://www.pvresources.com/en/top50pv.php>, (2008-1-17).
- [24] U.S. Department of Energy: *Solar Glossary of Terms*, Internet:
http://www1.eere.energy.gov/solar/solar_glossary.html,
(2008-1-17).
- [25] Wikipedia: *World energy resources and consumption*, Internet:
http://en.wikipedia.org/wiki/World_energy_resources_and_consumption,
(2006-02-06).
- [26] Wikipedia: *Photovoltaics*, Internet:
<http://en.wikipedia.org/wiki/Photovoltaic>, (2006-02-06).