



UPPSALA  
UNIVERSITET

TVE-ST5; 19002

Examensarbete 15 hp  
Juni 2019

# Single and multiple step forecasting of solar power production: applying and evaluating potential models

---

Hugo Uppling  
Adam Eriksson



UPPSALA  
UNIVERSITET

**Teknisk- naturvetenskaplig fakultet  
UTH-enheten**

Besöksadress:  
Ångströmlaboratoriet  
Lägerhyddsvägen 1  
Hus 4, Plan 0

Postadress:  
Box 536  
751 21 Uppsala

Telefon:  
018 – 471 30 03

Telefax:  
018 – 471 30 00

Hemsida:  
<http://www.teknat.uu.se/student>

## Abstract

### **Single and multiple step forecasting of solar power production: applying and evaluating potential models**

*Hugo Uppling & Adam Eriksson*

The aim of this thesis is to apply and evaluate potential forecasting models for solar power production, based on data from a photovoltaic facility in Sala, Sweden. The thesis evaluates single step forecasting models as well as multiple step forecasting models, where the three compared models for single step forecasting are persistence, autoregressive integrated moving average (ARIMA) and ARIMAX. ARIMAX is an ARIMA model that also takes exogenous predictors in consideration. In this thesis the evaluated exogenous predictor is wind speed. The two compared multiple step models are multiple step persistence and the Gaussian process (GP). Root mean squared error (RMSE) is used as the measurement of evaluation and thus determining the accuracy of the models. Results show that the ARIMAX models performed most accurate in every simulation of the single step models implementation, which implies that adding the exogenous predictor wind speed increases the accuracy. However, the accuracy only increased by 0.04% at most, which is determined as a minimal amount. Moreover, the results show that the GP model was 3% more accurate than the multiple step persistence; however, the GP model could be further developed by adding more training data or exogenous variables to the model.

Handledare: Linus Frosteryd & Victor Ingman  
Ämnesgranskare: Mahmoud Shepero  
Examinator: Joakim Widén  
ISSN: 1650-8319, TVE-ST; 19002

# Table of contents

<b>1. Introduction</b>	<b>2</b>
1.1 Aims	2
1.2 Research questions	3
1.3 Delimitations and limitations	3
1.3.1 Delimitations	3
1.3.2 Limitations	3
1.4 Thesis overview	3
<b>2. Background</b>	<b>4</b>
2.1 Photovoltaic production	4
2.2 Introduction to forecasting	4
2.3 Forecasting methods	5
2.3.1 Persistence	5
2.3.2 ARIMA	5
2.3.3 ARIMAX	6
2.3.4 Gaussian process	6
<b>3. Methodology</b>	<b>7</b>
3.1 Compared models	7
3.1.1 Persistence model	7
3.1.2 ARIMA model	7
3.1.3 ARIMAX model	7
3.1.4 Gaussian process model	7
3.2 Data	8
3.2.1 Data from the Sala facility	8
3.2.2 Data preprocessing for single step models implementation	8
3.2.3 Data preprocessing for multiple step models implementation	10
3.3 Model implementations	11
3.3.1 Single step models	11
3.3.2 Multiple step models	12
<b>4. Results</b>	<b>13</b>
4.1 Single step models	13
4.2 Multiple step models	13
<b>5. Discussion</b>	<b>16</b>
5.1 Single step models	16
5.2 Multiple step models	16
5.3 Summarizing discussion	17
<b>6. Conclusions</b>	<b>18</b>
<b>References</b>	<b>19</b>

# 1. Introduction

The European Union has committed to lower the greenhouse gas emissions by 80% to 95% between the years 1990 to 2050 (European Commission, 2012). Moreover, the European Commission has proposed solutions such as increase in the use of electricity production, renewable sources and system efficiency (European Commission, 2012).

In Sweden, the Swedish Energy Commission has set a goal for the electricity consumption to be supplied by 100% renewables before 2040 (Swedish Energy Commission, 2016), which will change the energy mix of today (van der Meer, 2018). In 2016, the Swedish Energy Agency presented their development strategy of photovoltaic (PV) electricity, stating that up to 10% of the total electricity consumed in 2040 could originate from PV production (Swedish Energy Agency, 2016).

Unlike traditional power generation based on fossil or nuclear fuels, power output from PV facilities could be considered highly variable because of its seasonal nature and daily irregularities due to cloudiness (Dughir et al., 2018). With this in mind, the expansion of electricity production from PV comes with certain challenges. Although solar energy is the most abundant energy resource, penetrations of PV power is connected with problems due to variability of power output, something which results in - among others - voltage fluctuations (Kleissl, 2013).

Solar power forecasting helps overcome the aforementioned challenges (Paulescu et al., 2018). Accurate forecasts for different time horizons act as a great tool in optimizing decision making and regulation of the grid (Kleissl, 2013). An important challenge is to communicate these forecasts in a format that provides enough information for system operators to gain decision support even though the forecasts are not 100% accurate (Letendre et al., 2014). As of 2013, the use of solar power forecasting within local PV facilities was not adopted by the market (Kleissl, 2013).

This thesis focuses on evaluating different potential forecasting models and model characteristics in an attempt to find accurate forecasting tools. The project is done in collaboration with Sun Labs, a company working mainly on data visualization and real-time data logging of solar cell facility production. Sun Labs collects data, such as energy, power, wind speed and pressure, from different PV installation sites in Sweden. The company wants to use their collected data to forecast PV power production.

## 1.1 Aims

The thesis aims to apply and evaluate potential forecasting models for solar power production, using the data measured from one of the PV installations of Sun Labs. More specifically, the thesis aims to compare the impacts of using the exogenous\* predictor wind speed in single-step predictions and also evaluate two multiple step prediction models for a 7 day forecast.

---

\* *Exogenous*: “something which affect the system but are not themselves affected” (Chatfield, 2016)

## 1.2 Research questions

To reflect the aims, the following research questions will be discussed in this thesis:

- Which of the various compared models has the highest single-step forecasting accuracy?
- How does the usage of an exogenous predictor change the performance of a single step forecasting models and ultimately the forecast accuracy?
- Which of the two compared models has the highest multiple step forecasting accuracy?

The compared models are presented in section 3.1.

## 1.3 Delimitations and limitations

### 1.3.1 Delimitations

This thesis will focus on a PV facility in Sala, Sweden. Therefore, the properties of the models will in this sense be site specific, meaning that the forecasting models will only be suitable for forecasts within this area. Further studies might investigate the generalizability of the obtained results.

### 1.3.2 Limitations

One important aspect of forecasting is the forecasting horizon. For accurate forecasting, the model needs historical training data (Kleissl, 2013). For a high-resolution dataset, the forecasting procedure is time consuming which limited the choice of models, model orders, training data and forecast time horizons. Furthermore, the dataset was inconsistent which further limited the scope of this study. Based on this, only data measured in July 2018 will be used. Additionally, the thesis will solely use wind speed as the exogenous predictor for the extended single step forecasting model.

In addition, data measurements stopped when the PV production was zero, i.e., during night hours, therefore wind speed was not recorded at these timestamps. The data will be described more in detail in section 3.2.

## 1.4 Thesis overview

The report firstly gives a short introduction to PV production and the various forecasting models in section 2. Section 3 contains the methodology, divided into explaining the dataset and the different evaluated models. This is followed by section 4, presenting the obtained results. Finally, the results are discussed in section 5 and conclusions are drawn in section 6.

To answer the defined research question, single and multiple step models will be implemented. The single and multiple step model implementations have separate sections for models, data preprocessing, implementations and results. The thesis follows the order of first presenting the single, then the multiple step models implementation.

## 2. Background

### 2.1 Photovoltaic production

PV power production converts sunlight directly into electricity (EIA, 2019). A PV cell consists of semiconducting materials which are affected by the sunlight such that when enough photons are absorbed by the material electrons start migrating towards one side of the cell, creating a voltage potential difference (EIA, 2019). The cells generate direct current electricity which is often converted into alternating current through an inverter (EIA, 2019).

Globally, almost 100 GW of PV was installed during 2017, making it the fastest growing energy technology, seen in regards to net installed capacity (IEA, 2017). The biggest contributor, representing more than 50% of the newly installed PV, is China, followed by the USA and India (IEA, 2017). In comparison with other sources of electricity, PV accounted for 2% of the total amount of generated electricity in 2017 (IEA, 2017).

In Sweden, installations of PV power systems continue to increase. From the year 2016 to 2017, 117.6 MW new capacity PV power were installed, representing a growth of more than 50% compared to the year before (Lindahl and Stoltz, 2017). The Swedish PV market is divided into commercial and privately owned systems, where 62% of the installations are commercial systems and 34% are installed by private entities (Lindahl and Stoltz, 2017). Both markets consist of almost exclusively roof-mounted systems (Lindahl and Stoltz, 2017). The remaining 4% of the market consists of small PV solar parks (Lindahl and Stoltz, 2017). In total, the installed amount of grid-connected PV power 2017 was estimated to be around 307 MW, which is expected to increase in the coming years (Lindahl and Stoltz, 2017).

In order to reach the targets set by the Swedish Energy Commission, the Swedish government has deployed both subsidies and incentives to encourage investing into PV systems. The PV market has through direct subsidies received a total of one billion SEK at the end of 2017 (Lindahl and Stoltz, 2017).

### 2.2 Introduction to forecasting

For many decades, grid load forecasting has been a central part of managing the electricity grid markets related to electricity production (Kleissl, 2013). Since electricity from solar power is set to be a more important source of energy, the importance of forecasting solar power production will continue to increase (van der Meer, 2018).

There is a wide range of methods for solar power forecasting. Models can be based on either physical or stochastic correlations (Kleissl, 2013). Physical models often use images from satellites for tracking and predicting locations of clouds, while stochastic models are based on time series forecasting (Kleissl, 2013). Which method to apply depends on different factors such as accessibility of data and the time horizon of the forecast (Kleissl, 2013). Time horizons can be divided into three groups: Nowcasting, short to medium-term forecasting, and long-term forecasting (Hu et al., 2015).

Forecasting with time horizons ranging from a few seconds to a few hours is called nowcasting (Dughir et al., 2018). The length of the time horizon depends on the resolution of the dataset and the aim of the forecast itself (Dughir et al., 2018). These forecasting methods have historically been used by the grid operators to match supply and demand as well as reducing the need for regulation in real time (Kleissl, 2013). Nowcasting of solar power production will be of increasing importance in a smart grid environment (Hu et al., 2015).

When forecasting for a time horizon up to 7 days, short to mid-term forecasting, the area of value creation is within energy resource scheduling and planning (Kleissl, 2013). The shorter forecasts within this time horizon could be used in deciding for example unit commitment, while the longer ones are used when planning maintenance of different components in the solar power system (Hu et al., 2015).

Long-term forecasting treats predictions for months up to a year ahead and are used for example in planning new PV plants or when negotiating with financing or distribution entities (Hu et al., 2015).

## 2.3 Forecasting methods

Different models have been used in nowcasting and short to mid-term solar power forecasting; examples of models applied are Support Vector Machines, Takagi-Sugeno fuzzy model, ARIMA (Qiao and Zeng, 2013), neural networks and Gaussian process (Bonilla and Dahl, 2017). The models applied in this thesis are persistence, ARIMA, ARIMAX and Gaussian process, and will be presented in the following sections.

### 2.3.1 Persistence

The persistence method is the simplest way of forecasting time series (Kleissl, 2013). This model is based upon the assumption that conditions stays constant and therefore the next predicted value will be the same as the previous time-step (Kleissl, 2013). Because of the characteristics of solar power production, the persistence model will not perform accurate for medium or longer forecasts (Kleissl, 2013). However, given a high-resolution dataset, the persistence forecast for shorter periods, intra hour, often performs as good as more advanced models (Kleissl, 2013).

### 2.3.2 ARIMA

The second model applied in the thesis is the autoregressive integrated moving average (ARIMA) model (Hyndman, 2018). This model is commonly used in analyzing non-stationary processes and time series where values do not fluctuate around a constant mean (Kleissl, 2013). As the name implies, this model consists of three main characteristics. The autoregressive (AR) component provides the regression part where the forecast uses a linear combination of earlier values of the variable against itself (Hyndman, 2018). ARIMA makes the dataset stationary through differencing which is conversely represented in the integrated (I) component of the model (Hyndman, 2018). Finally, the moving average component (MA) uses earlier forecast errors to further improve the model (Hyndman, 2018).

The three components create the order (p, d, q) of the ARIMA model, where p controls autoregression, i.e., how far back to regress, d controls the order of differencing, and q controls the number of lags of errors taken into consideration by the model (Kleissl, 2013).

### **2.3.3 ARIMAX**

An ARIMAX model is an ARIMA model extended with an exogenous variable (Gooi and Sanjari, 2016). This forecast methodology is useful when a dataset is closely related to other sampled data or dependent on other variables (Pomares et al., 2017).

### **2.3.4 Gaussian process**

The Gaussian process (GP) model was implemented as a multiple step forecast. A GP is a collection of random variables, any finite number of which have a joint Gaussian distribution (Rasmussen, 2006). The GP is defined by the kernel (or covariance) function together with the mean function (Rasmussen, 2006). Given a set of training data and assumptions about the kernel characteristics, the GP defines a function which returns a value to any possible input value (Rasmussen, 2006). A future input time step will return a forecasted value at the input time step. The assumptions of the characteristics of the kernel are represented in the kernel function of the GP and will be described in section 3.1.4. There are multiple kernel functions to apply for a GP model; linear, periodic and squared-exponential kernel functions are most commonly used (Duvenaud, 2014). Kernels are used and combined to create a model function that fits the model data (Duvenaud, 2014).

The GP provides a nonparametric approach to regression (Bonilla and Dahl, 2017). Despite the GP model's computational complexity  $\mathcal{O}(N^3)$ , where N is the number of training values, the choice of kernel function, or assumptions about the kernel characteristics, gives flexibility to the nonlinear model (Duvenaud, 2014). The GP has been used in forecasting solar power (Sheng et al., 2017), wind power (Yan et al., 2016), electric load (Shepero et al., 2018), and electricity pricing (Alamaniotis et al., 2015).



## 3. Methodology

The methodology section is divided into three main subsections. Section 3.1 describes the mathematical foundations for the various models evaluated in the thesis. Section 3.2 describes the dataset and how it has been processed. Lastly, section 3.3 explains the two different model implementations.

### 3.1 Compared models

This section will present the different mathematical formulations of the four previously presented models, following the same order as section 2.3.

#### 3.1.1 Persistence model

The persistence model is given by:

$$\hat{y}_t = y_{t-1}, \quad (1)$$

where the forecasted value  $\hat{y}_t$  is given by the previous value  $y_{t-1}$  (Kleissl, 2013).

#### 3.1.2 ARIMA model

For the differenced time series, the ARIMA(p,d,q) model is described as:

$$y'_t = c + \varphi_1 y'_{t-1} + \dots + \varphi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} + \varepsilon_t. \quad (2)$$

The time series value  $y'_t$  is given by a constant plus a linear combination of  $p$  lags of earlier values plus a linear combination of  $q$  lags of forecasted errors (Hyndman, 2018). The  $d$  in the model order is the order of differencing of  $y'_t$  (Hyndman, 2018).

#### 3.1.3 ARIMAX model

The ARIMAX equation for the stationary time series value  $y'_t$  is described as:

$$y'_t = bx_t + \varphi_1 y'_{t-1} + \dots + \varphi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} + \varepsilon_t, \quad (3)$$

which is similar to Equation 2 with the exception of the first term,  $bx_t$ . The  $x_t$  represents the exogenous predictor at time  $t$  and  $b$  is its coefficient (Shumway and Stoffer, 2017).

#### 3.1.4 Gaussian process model

The GP is given by the mean function  $m(x)$  and covariance function  $k(x, x')$  of an actual process  $f(x)$  (Rasmussen, 2006):

$$m(x) = \mathbb{E}[f(x)], \quad (4)$$

$$k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))]. \quad (5)$$

The GP is defined as (Rasmussen, 2006):

$$f(x) \sim GP(m(x), k(x, x')). \quad (6)$$

This thesis combines a linear kernel function  $k_{lin}$  and a periodic kernel function  $k_{per}$ , a method pursued by Bonilla and Dahl (2017). The linear kernel function is described as:

$$k_{lin}(x, x') = \sigma_f^2(x - c)(x' - c), \quad (7)$$

where  $\sigma_f^2$  is a scale factor and  $c$  are offsets of  $x$  and  $x'$  (Duvenaud, 2014). The periodic kernel is given by:

$$k_{per}(x, x') = \sigma_f^2 \exp\left(-\frac{2}{l^2} \sin^2\left(\pi \frac{x-x'}{p}\right)\right), \quad (8)$$

The variable  $\sigma_f^2$  is the scale factor,  $l$  is the lengthscale and  $p$  is the period (Duvenaud, 2014).

The overall kernel function  $k(x, x')$  is the two kernels multiplied with each other:

$$k(x, x') = k_{lin} \cdot k_{per}. \quad (9)$$

## 3.2 Data

This section is divided in three parts. First, the Sala facility data will be described in section 3.2.1. Then the report will walk through the steps of preprocessing production and wind speed data for single step models implementation in section 3.2.2. Lastly, the data preprocessing of the multiple step models implementation will be presented in section 3.2.3.

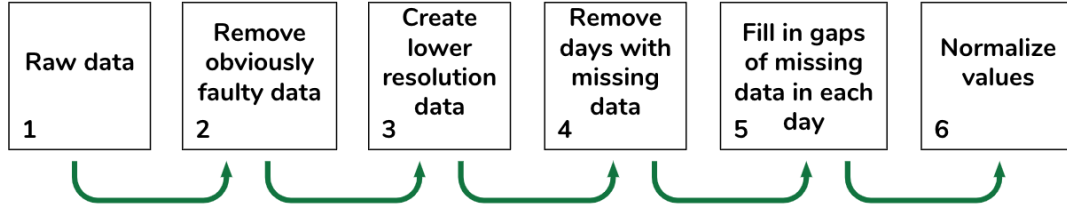
### 3.2.1 Data from the Sala facility

The facility is placed in Sala, Sweden and was installed in late 2009 (Solel i Sala & Heby ekonomiska förening, n.d.). The site has an installed maximum power of 47 kW and a yearly energy production of 40 MWh (Solel i Sala & Heby ekonomiska förening, n.d.). Sun Labs started collecting data from the Sala facility in May 2018. The complete dataset of the Sala facility ranged from mid-May 2018 to mid-January 2019, and contained multiple measurements such as instantaneous power production [W], energy produced per day [J], wind speed [m/s], illuminance [lx], and ambient temperature [°C]. From this dataset, production and wind speed in the month of July 2018 were used, as this month covered the most consistent sets of production and wind speed data. In the thesis, the non-exogenous variable is the PV production, and the exogenous variable is wind speed.

### 3.2.2 Data preprocessing for single step models implementation

To implement the single step models, data from both the production as well as the exogenous predictor wind speed were required. Firstly, the production data, and then the wind speed data were preprocessed. This section will follow this chronological order.

The single step models used raw data gathered from July 2018, which has been preprocessed as described in the flowchart presented in Figure 1. The raw data was processed to create two different time steps of 5 minutes and 10 minutes. This step (step 3 in Figure 1 as well as in Figure 2) will be explained more in detail in the single step models implementation in section 3.3.1.



*Figure 1. Flowchart of production data preprocessing for the month of July for the single step models implementation.*

Obviously faulty data such as outlier timestamps or single outlier production values were first removed (step 2 in Figure 1). The interval between consecutive recorded values was, for 97% of the data, between 0 to 30 seconds. The remaining intervals were larger than 30 seconds and occurred mostly when production stopped and the instrument did not record values until production started the next day. To create lower resolution data that is uniformly sampled with either 5- or 10-minute intervals, mean values for every time steps were created (step 3 in Figure 1). This was done by filling the inter-sample values with the preceding value for every second, and then calculating mean values for every 5 or 10 minutes.

The days of July 22nd, 30th and 31st were removed because of the scarcity of the production data measured at the given days (step 4 in Figure 1). This is because of connectivity issues that occurred between the facility instrument and the company database. The final dataset therefore represented 28 days in July 2018.

As mentioned in section 1.3.2, the instrument stopped measuring when production stopped. Therefore, the values between production end and start were assumed and filled in as zeros. Furthermore, gaps in the dataset at certain time steps during the day were filled in with the monthly average value of that time step (step 5 in Figure 1). Finally, the production values were normalized to the maximal capacity of the month, to create a relative percentage error, according to Hoff et al. (2013):

$$P_n = \frac{p}{p_{max}}, \quad (10)$$

where  $p_n$  is the normalized value,  $p$  [W] is the production value and  $p_{max}$  [W] is the maximum production value of the month.

After the production was preprocessed, the wind speed data of July was preprocessed in a similar manner, see the flowchart shown in Figure 2.

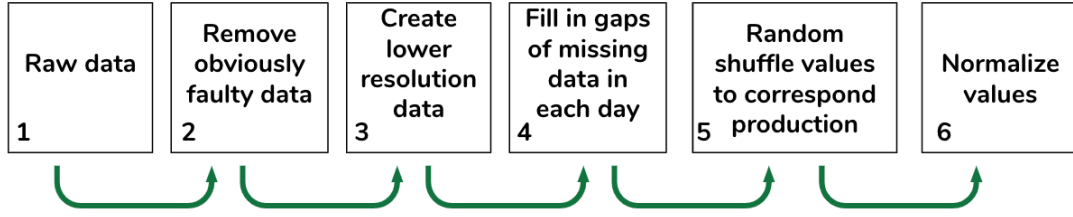


Figure 2. Flowchart of wind speed data preprocessing for the month of July for the single step models implementation.

First, obviously faulty data such as extreme wind speeds (over 32 m/s) and outlier timestamps, such as dates in the year 2030, were removed (step 2 in Figure 2). After this, lower resolution data was created (step 3 in Figure 2). Like the production dataset, gaps in the wind speed dataset were filled in with the monthly average wind speed value at that time step (step 4 in Figure 2). Finally, the wind speed values were matched to the production data, in the regards to having the same days in the same order (step 5 in Figure 2). The final normalized values of wind speed according to equation 10 (step 6 in Figure 2) are the same length as the corresponding time step of the production data. The size of the four datasets are presented in Table 1.

Table 1. The length of a day and the total values of July of wind speed and production at the different time steps 5 or 10 minutes.

<i>Production and wind speed values</i>	<b>5 minutes</b>	<b>10 minutes</b>
<b>Values per day</b>	288	144
<b>Total values</b>	8064	4032

### 3.2.3 Data preprocessing for multiple step models implementation

The implemented multiple step models used the production data of July for training and testing, explained in section 3.3.2. This required a similar preprocessing process to the one presented in section 3.2.2, with some adjustments. A flowchart visualizes the process in Figure 3.

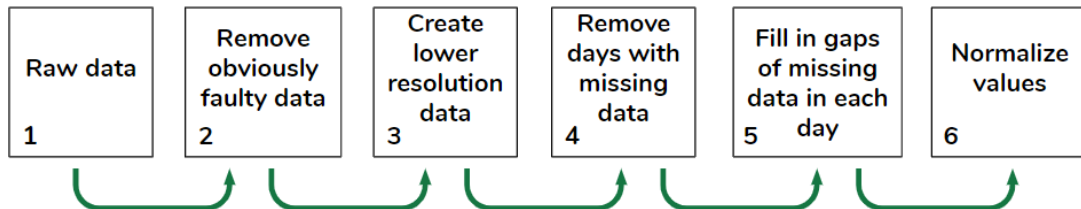


Figure 3. Flowchart of production data preprocessing for the month of July for the multiple step models implementation.

Analogously, the same obviously faulty data was removed as the preprocessed production data for single step model implementations (step 2 in Figure 3). After that, a

lower resolution of 5-minute time steps was created (step 3 in Figure 3). The days 22nd, 30th and 31st of July were removed due to the days having missing data (step 4 in Figure 3). The dataset of July for the multiple step models also contains 28 days. Precedingly, missing data in the 28 days were filled in using the average of that particular time step in July (step 5 in Figure 3). Finally, the production was normalized according to equation 10 (step 6 in Figure 3).

### 3.3 Model implementations

To be able to answer the set research questions the five models (persistence, ARIMA, ARIMAX, GP and multiple step persistence) will be implemented into two separate implementations: single step models, and multiple step models. The model implementations will be presented in section 3.3.1 for single step models, and in section 3.3.2 for the multiple step models.

The evaluation measurement of both the single step and the multiple step models is the respective root mean squared error (RMSE) as suggested by (Hoff et al., 2013):

$$RMSE = \sqrt{\frac{\sum_{n=1}^N (\hat{y}_n - y_n)^2}{N}}, \quad (11)$$

where  $N$  is the number of predictions,  $\hat{y}$  is the prediction and  $y$  is the actual value.

#### 3.3.1 Single step models

The single step models implementation will evaluate the performance of single step prediction of the persistence, ARIMA and ARIMAX models. The ARIMAX models have the exogenous predictor normalized wind speed.

As a sensitivity analysis of the model order and data resolution, the models were simulated in the two different time steps: 5 minutes and 10 minutes. In addition, the ARIMA and ARIMAX models were simulated in orders (5,1,0), (5,1,1) and (4,1,2). These orders were decided by minimizing the AIC value - as stated by Hyndman (2018) - of the fitted models as well as ensuring that the residuals\* were uncorrelated and normally distributed, following the procedure of ARIMA model evaluation according to Hyndman (2018).

The ARIMA and ARIMAX models were trained using the first 24 days ( $\approx 86\%$  of dataset) of July after randomly shuffling the days, and are summarized in the Table 2. The days were randomly shuffled within the months to decrease the impact of the weather seasonality over multiple days. The models ARIMA, ARIMAX and persistence forecast the next time step for the remaining four days ( $\approx 14\%$ ) of the July dataset. After forecasting a single time step, the models were refitted using the new observations to forecast the next time step. This means that the models predicted on 1152 and 576 steps for the time steps 5- and 10-minutes respectively.

---

\* Residual is the forecasts deviation from the actual value (Hyndman, 2018)

Table 2. The number of values of historic and forecasting data for the single step models.

<i>Production and wind speed values</i>	<b>5 minutes</b>	<b>10 minutes</b>
<b>Values per day</b>	288	144
<b>History</b>	6912	3456
<b>Number of forecasts</b>	1152	576

### 3.3.2 Multiple step models

The multiple step models implementation will compare the performance of the GP forecast to a multiple step persistence forecast, i.e., forecasting the production of the week to be the same as the previous week at each corresponding time step.

The evaluated GP model used the kernel  $k$ , consisting of a linear kernel  $k_{lin}$  and a periodic kernel  $k_{per}$ , as described in section 3.1.4.

The GP model was trained using history of the first 21 days (75% of the dataset) of July and forecasted on the last 7 days (25% of the dataset) of July. The training data for each day excluded the night times where zero production was recorded, this approach was pursued by Bonilla and Dahl (2017). In other words, only production values between 01:30 and 20:30 are considered, creating a forecast day of 19 hours. The time step was 5 minutes which created 229 data points per day. The kernel function was learned using the *GaussianProcessRegressor* function of the scikit-learn library (Scikit-learn, n.d.).

The multiple step persistence forecasted the last 7 days of July as the previous 7 days.

Table 3. The number of values of historic and forecasting data for the multiple step model.

<i>Production values</i>	<b>5 minutes</b>
<b>Values per day</b>	229
<b>History</b>	4809
<b>Forecast</b>	1603

## 4. Results

In section 4.1, the results of the single step models implementation are presented and in section 4.2 the results of the multiple step models implementation are presented.

### 4.1 Single step models

Fourteen models are presented in total; there were six ARIMA and ARIMAX models each, with orders (5,1,0), (5,1,1) and (4,1,2) and time steps 5 minutes and 10 minutes, as well as two persistence models for the respective time steps. The RMSE of each simulated model over four days of single step forecasting is presented in Table 4.

The single step ARIMA and ARIMAX models were marginally better than the persistence model. The biggest improvement of an ARIMA or ARIMAX model compared to persistence, was the ARIMAX(4,1,2) model, which had an RMSE 0.23% lower than the persistence model of the 10-minute time step.

The ARIMAX models were consistently better than both the ARIMA, improvement ranging between 0.02% and 0.04%, and the persistence model, improvement ranging between 0.13% and 0.23%.

*Table 4. The RMSE of each single step forecasting model. To make comparison easier, the RMSE of the two persistence models, of the time steps 5 and 10 minutes, are presented in each ARIMA and ARIMAX model order. ARIMAX has the lowest RMSE for every model order and time step.*

RMSE	(5,1,0)		(5,1,1)		(4,1,2)	
	5 minutes	10 minutes	5 minutes	10 minutes	5 minutes	10 minutes
<b>Persistence</b>	7.65%	8.89%	7.65%	8.89%	7.65%	8.89%
<b>ARIMA</b>	7.49%	8.72%	7.52%	8.73%	7.54%	8.69%
<b>ARIMAX</b>	7.47%	8.68%	7.49%	8.70%	7.52%	8.66%

### 4.2 Multiple step models

The multiple step models implementation provides a multiple step persistence forecast mirroring the week before and a GP forecast providing a weekly forecast based on the 21 days of training history, as presented in section 3.3.2. The RMSE of the GP model and the persistence model are presented in Table 5. The GP model had the lower RMSE of 19.6% of the normalized production whereas the persistence model has a RMSE of 22.6%. The GP model had a 3% lower RMSE compared to the multiple step persistence model.

Table 5. The RMSE of forecasting seven days with the GP and multiple step persistence models with time step of 5 minutes, with 21 days of history.

<i>Model/Error</i>	<b>RMSE of normalized production</b>
<b>Persistence</b>	22.61%
<b>Gaussian process</b>	19.64%

After observing the GP forecast, presented in Figure 4, the authors noticed that the last two days of the week were especially cloudy days. When forecasting the first five days, instead of seven, presented in Figure 5, the RMSE of the GP model improved to 13.07%, meanwhile the persistence forecast had an RMSE of 20% for the same five days. Thus, without the two cloudy days the RMSE decreased by 6.6% for the GP model and by 2.6% for the multiple step persistence model. This indicates that a longer training dataset might be useful in improving the fit of the GP model.

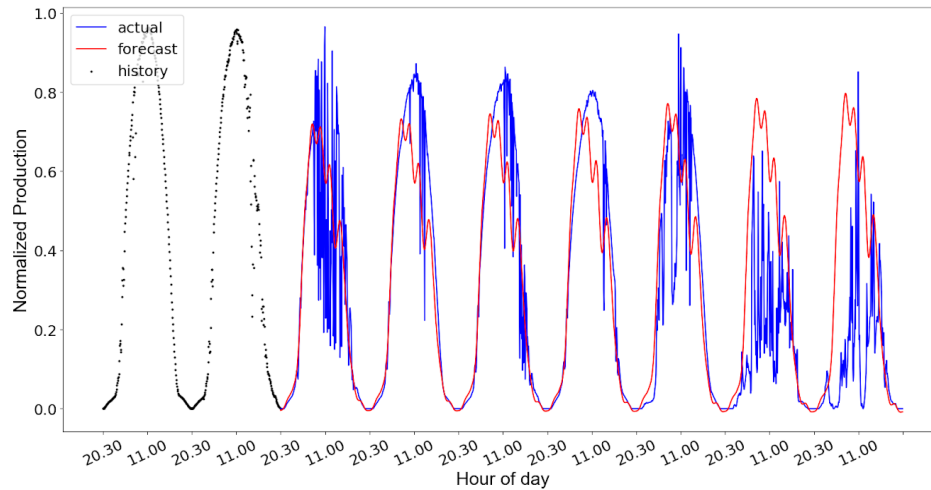
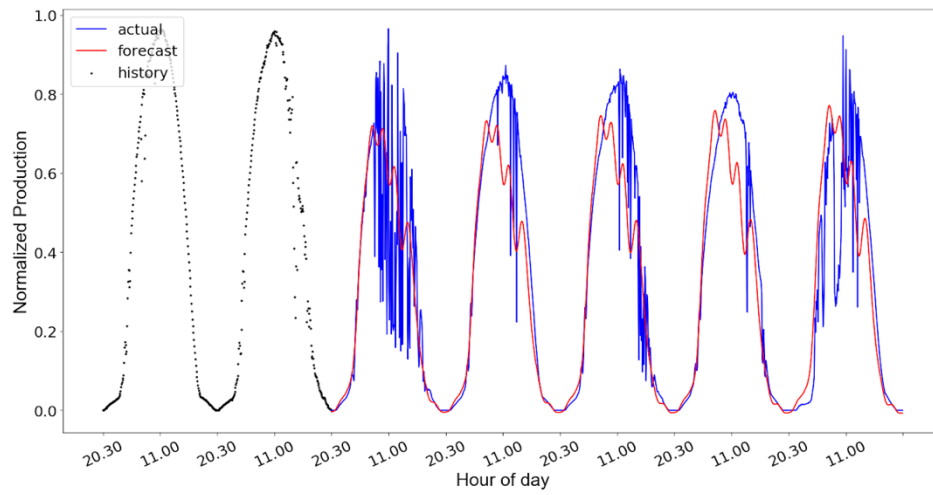


Figure 4. Plot of 2 out of 21 days history (black dots) and the actual values (blue) and GP forecast (red) for 7 days. The values from 20:30 to 01:30 were excluded from this plot, see the explanation in section 3.3.2.





*Figure 5. Plot of 2 out of 21 days history (black dots) and the actual values (blue) and GP forecast (red) for 5 days. The values from 20:30 to 01:30 were excluded from this plot, see the explanation in section 3.3.2.*

## 5. Discussion

The discussion firstly evaluates and reflects on the results of the single step models implementation and the multiple step models implementation in sections 5.1 and 5.2 respectively. After this, a broader discussion is provided to find a common ground between the two implementations.

### 5.1 Single step models

The errors from the four days of prediction from the ARIMA and ARIMAX models with 24 days of training history were, as shown in Table 4, close to the errors of the persistence model. The ARIMA and ARIMAX models were more accurate than the persistence model for every time step and model order. A conclusion to draw is that both the ARIMA and ARIMAX models are more accurate than the persistence model when performing single step forecasting for four days of July with the time resolution 5 and 10 minutes. Further, the results showed that the ARIMAX models are consistently more accurate than the ARIMA models, in every order and time resolution.

However, the accuracy of either the ARIMA or the ARIMAX model had not improved by much in any case. The errors decreased by at most 0.23% when compared to the persistence model. The simulated ARIMA and ARIMAX models take long computational time since the linear models need to be fitted to the training data, whereas the persistence model takes no computational time. Further, a model order has to be chosen, which is time consuming compared to solely using the persistence model.

For both the ARIMA and ARIMAX models, the order (4,1,2) had the lowest RMSE of 8.69% and 8.66% respectively, in the 10-minute time step. Further, in the 5-minute time step the order (5,1,0) had the lowest RMSE of 7.49% and 7.47% respectively. When only observing the ARIMAX and ARIMA RMSE for every order presented in the result with a 10-minute time step, one can notice that the error first increased between (5,1,0) and (5,1,1), and then decreased for the order (4,1,2). When observing the RMSE of the ARIMA and ARIMAX models in the 5-minute time step, the RMSE increased from (5,1,0), to (5,1,1) and (4,1,2). Even though increasing the moving average parameter  $q$  gave a better fit, in regards to AIC and the residual analysis, it does not mean that the forecast accuracy is better. This conclusion can be drawn with all presented ARIMA and ARIMAX orders, as none of models outperformed the persistence model by more than 0.2% and is in line with Hyndman's statement that a "*model which fits the training data well will not necessarily forecast well.*" (Hyndman, 2018, ch:3.4)

### 5.2 Multiple step models

As described in section 3.1.4, the GP model employed a kernel consisting of linear and periodic kernel functions. The characteristic of the defined kernel is a periodic function with an increasing amplitude as given by (Duvenaud, 2014). In the Figures 4 and 5, the GP forecasts of the last 5 and 7 days of the July dataset were presented. The forecast function has a clear periodic daily trend. The increasing amplitude trend was not as visible in either the 7 or 5 days forecast. This was desirable, as solar power production does follow the kernel trend of continuously increasing amplitude. The choice of coefficients in the linear kernel were made to reflect this fact. However, for longer

forecasts with the learned GP model, the rising amplitude trend might be more apparent. To fit a longer forecast horizon, a different kernel function might be better employed.

The forecast, shown in Figure 4, is neither of the same amplitude as the first 5 (non-cloudy) days of the forecast nor the last 2 (cloudy) days of the forecast. This is due to the amplitude represented the average of the cloudy and non-cloudy days from the training data. If the GP model were to train on 21 non-cloudy days, the forecast function would have given a more accurate forecast in the first 5 days, compared with the forecast of the GP model presented in the results. However, if the actual week of the forecast were cloudy, the GP forecast error would increase. The GP model was trained on both cloudy and non-cloudy days, and could not identify the difference, and therefore forecasted a result in between the two types of days.

The results indicate that the model does not have enough history training data. However, there could be other solutions too. Including exogenous variables in a model can help identify production trends and forecast, as suggested in Qiao and Zeng (2013). Another solution is to use other kernel characteristics to identify other periodicity than the daily trends. Unfortunately, it was not in the scope of this thesis to further develop non-linear models with exogenous variables, different kernel structures or longer training history, and remains a point to look at in future work.

### 5.3 Summarizing discussion

This thesis evaluates the performance of various potential forecasting models through answering the research questions in section 1.3. According to the results, the most accurate single step forecasting model is the ARIMAX, independent of the model order (p,d,q) which entails that the use of the exogenous variable wind speed makes the forecast more precise. However, it is important to be careful in drawing extensive conclusions out of these results since the positive correlation between the two variables, is considerably small (at max a 0.04% decrease in RMSE). Because of computational time and model fitting, one could argue that the persistence model is just as good as or even better than the ARIMA and ARIMAX models presented in the results.

In the evaluation of the multiple step forecasting models, the GP model is more accurate than the persistence model by 3%. However, the 5 and 7 day forecast horizons show that the GP model could be further developed, with a longer training history, different kernel characteristics or adding exogenous variables in the non-linear multiple step forecast model, as discussed in section 5.2.

For both model implementations, common limitations are computational time and data availability. As the results indicate in section 5.2, the GP model would have a more accurate forecast with a longer training dataset. Further, multiple ARIMA and ARIMAX model orders could be applied with a longer forecast length, and chosen from the lowest RMSE instead of minimizing the AIC. If the scope of the thesis or the availability of data would allow it, more developed models could have been evaluated. It would be of interest to evaluate how adding another exogenous predictor to the ARIMAX models would change its performance, or to evaluate another predictor with possibly more impact than wind speed. Likewise, evaluating a non-linear forecasting model with exogenous predictors would be of interest to decide the impact of having more data and training history.

## 6. Conclusions

This thesis seeks to find the highest single step model forecasting accuracy, comparing three ARIMA and ARIMAX orders with one persistence model, for the two different time step resolutions. The ARIMAX model is the most accurate for every model order and time step; however, the model with the exogenous variable wind speed is minimally more accurate than the ARIMA or the persistence model. The report has shown that the usage of the exogenous predictor wind speed does improve the accuracy in the single step models implementation, although by at most 0.04%, a negligible amount.

The thesis also had the ambition to find the most accurate multiple step forecasting model, comparing the GP with a multiple step persistence model. The results showed that the implemented GP model is 3% more accurate than a multiple step persistence model. However, the results also show that the GP model could be further developed to increase the forecast accuracy.

Future work might include a more thoroughly investigation of ARIMA and ARIMAX model orders in the single step models, and further develop the GP model according to suggestions in section 5.3, to be able to create more accurate forecasts for solar power production.

## References

- Alamaniotis, M., Bourbakis, N., Tsoukalas, L.H. (2015), Very-short term forecasting of electricity price signals using a Pareto composition of kernel machines in smart power systems, 2015 IEEE Global Conference on Signal and Information Processing, Electronic ISBN: 978-1-4799-7591-4. Available online: <https://ieeexplore.ieee.org/document/7418303> (2019-05-28)
- Bonilla, E., Dahl, A. (2017) Scalable Gaussian Process Models for Solar Power Forecasting. In: Woon W., Aung Z., Kramer O., Madnick S. (eds) Data Analytics for Renewable Energy Integration: Informing the Generation and Distribution of Renewable Energy. DARE 2017. Lecture Notes in Computer Science, vol 10691. Springer, Cham
- Chatfield, C. (2016), The analysis of time series: An introduction, 6th edition, CRC Press LLC
- Dughir, C., Oana, M., Paulescu, E., Paulescu, M. (2018), Nowcasting the Output Power of PV Systems, EDP Sciences, Available online: [https://www.e3s-conferences.org/articles/e3sconf/abs/2018/36/e3sconf\\_icren2018\\_00010/e3sconf\\_icren2018\\_00010.html](https://www.e3s-conferences.org/articles/e3sconf/abs/2018/36/e3sconf_icren2018_00010/e3sconf_icren2018_00010.html) (2019-05-20)
- Duvenaud, D.K. (2014), Automatic Model Construction with Gaussian Processes, University of Cambridge, Available online: <https://www.cs.toronto.edu/~duvenaud/thesis.pdf> (2019-05-20).
- European Commission. (2012), Energy roadmap 2050. Available online: [https://ec.europa.eu/energy/sites/ener/files/documents/2012\\_energy\\_roadmap\\_2050\\_en\\_0.pdf](https://ec.europa.eu/energy/sites/ener/files/documents/2012_energy_roadmap_2050_en_0.pdf) (2019-05-21)
- Gooi, H. B., Sanjari M. J. (2016), Probabilistic Forecast of PV Power Generation Based on Higher Order Markov Chain, 16969918, IEEE, Available online: <https://ieeexplore.ieee.org/document/7592462> (2019-05-20)
- Hoff, T. E., Perez, R., Kleissl, J., Renne, D., Stein, J. (2013), Reporting of irradiance modeling relative prediction errors. Progress in Photovoltaics: Research and Applications, 21, Wiley Online Library Available online: <https://doi-org.ezproxy.its.uu.se/10.1002/pip.2225> (2019-05-28)
- Hu, Z., Lin, J., Song, Y., Wan, C., Xu, Z., Zhao, J. (2015), Photovoltaic and Solar Power Forecasting for Smart Grid Energy Management, CSEE Journal of Power and Energy Systems, Vol. 1, No 4. Available online: <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7377167> (2019-05-20).
- International Energy Agency (IEA). (2018), Annual Report 2017, International Energy Agency Available online: <http://www.iea-pvps.org/index.php?id=6> (2019-05-20)
- Kleissl, J. (2013), Solar Energy Forecasting and Resource Assessment, Academic Press.
- Letendre, S., Makhyoun, M., Taylor, M. (2014), Predicting Solar Power Production: Irradiance Forecasting Models, Applications and Future Prospects, Solar Electric Power Association and Green Mountain College Vermont, Washington DC
- Lindahl, J., Stoltz, C. (2018), National Survey Report of PV Power Applications in Sweden 2017, Swedish Energy Agency & International Energy Agency (IEA), Available online:

- [http://www.ieapvps.org/index.php?id=93&no\\_cache=1&tx\\_damfrontend\\_pi1%5BshowUid%5D=740&tx\\_damfrontend\\_pi1%5BbackPid%5D=93](http://www.ieapvps.org/index.php?id=93&no_cache=1&tx_damfrontend_pi1%5BshowUid%5D=740&tx_damfrontend_pi1%5BbackPid%5D=93) (2019-05-20).
- Pomares, H., Rojas, I., Valenzuela, O. (2017), Time Series Analysis and Forecasting Selected Contributions from ITISE 2017, Springer.
- Qiao, W., Zeng, J. (2013), Short-term solar power prediction using a support vector machine, Science direct.
- Rasmussen, C. E., Williams, K. I. (2006), Gaussian Processes for Machine Learning, The MIT Press Cambridge, Massachusetts
- Scikit-learn. (No date), Gaussian process regression. Available online: [https://scikit-learn.org/stable/modules/generated/sklearn.gaussian\\_process.GaussianProcessRegressor.html](https://scikit-learn.org/stable/modules/generated/sklearn.gaussian_process.GaussianProcessRegressor.html) (2019-05-21)
- Sheng, H., Xiao, J., Cheng, Y., Ni, Q., Wang, S. (2018) Short-term solar power forecasting based on weighted Gaussian process regression, IEEE Transactions on Industrial Electronics, Vol. 65, Issue 1, Pages 300-308. Available online: <https://ieeexplore.ieee.org/document/7945510> (2019-05-28).
- Shepero, M., van der Meer, D., Munkhammar, J., Widén, J. (2018), Residential probabilistic load forecasting: A method using Gaussian process designed for electric load data, Applied Energy, Vol. 218, Pages 159-172. Available online: <https://doi-org.ezproxy.its.uu.se/10.1016/j.apenergy.2018.02.165> (2019-05-28).
- Shumway, R. H., Stoffer, D. S. (2017), Time Series Analysis and Its Applications With R Examples, 4th edition, Springer
- Solel i Sala & Heby ekonomiska förening. (no date), Anläggningarna. Available online: <http://www.solelisaheby.se/anlaggningarna/> (2019-05-20)
- Swedish Energy Agency. (2016), Förslag till strategi för ökad användning av solel, Available online: [http://www.energimyndigheten.se/globalassets/fornybart/solenergi/solen-i-samhallet/forslag-till-strategi-for-okad-anvandning-av-solel\\_webb.pdf](http://www.energimyndigheten.se/globalassets/fornybart/solenergi/solen-i-samhallet/forslag-till-strategi-for-okad-anvandning-av-solel_webb.pdf) (2019-05-28).
- Swedish Energy Commission. (2016), Ramöverenskommelse mellan Socialdemokraterna, Moderaterna, Miljöpartiet de gröna, Centerpartiet och Kristdemokraterna, Available online: <https://www.regeringen.se/contentassets/b88f0d28eb0e48e39eb4411de2aabe76/energioverenskommelse-20160610.pdf> (2019-05-28).
- Van der Meer, D. (2018), Spatio-temporal probabilistic forecasting of solar power, electricity consumption and net load, Uppsala University.
- U.S. Energy Information Administration (EIA). (2019), Photovoltaics and Electricity. Available Online: [https://www.eia.gov/energyexplained/index.php?page=solar\\_photovoltaics](https://www.eia.gov/energyexplained/index.php?page=solar_photovoltaics) (2019-05-20)
- Yan, J., Li, K., Bai, E.W., Deng, J., Foley, A.M. (2016), Hybrid probabilistic wind power forecasting using temporally local Gaussian process, IEEE Transactions on Sustainable Energy, Vol. 7, Issue 1, Pages 87-95. Available online: <https://ieeexplore.ieee.org/document/7274766> (2019-05-28).