Analysis of Forecasts for District Heat Production Using Different Models for Seasonal Partitions

Leo Einarsson - π16

Abstract District heating is a common means of space and hot water heating in Sweden. However, the demand for heating is not the same at all times. On a yearly basis more heat is required during winter, while next to none is needed in summer. Since the demand for heat load varies throughout the year, when trying to predict it, using a model that changes with the seasons can give a more accurate prediction. In this study, a forecasting model was tested to change its parameters either yearly, every three months (seasonal), monthly or weekly. The goal was to see which way of partitioning the year would give a more reliable prediction. Using statistical bootstrap to create confidence and prediction bands for the heat load, an analysis was conducted. The results show that a seasonal or monthly approach give a more accurate prediction overall and that the summer was most difficult to predict, relative to the produced heat, although transition seasons, for instance between spring and summer were more prone to large variances overall.

Preface

With this thesis work, Leo Einarsson will finish their studies at the Engineering Faculty in Lund's programme in Engineering Mathematics. This project has been performed in cooperation with a company dealing with with district heating forecasts during springtime 2021.

I'd especially like to thank Björn Kjeang Funkqvist, my supervisor at the company, as well as Dragi Anevski and Maria Sandsten, my supervisor and examinator respectively at the department of mathematical statistics.

Another big Thank you to everyone at the company who have taken their time to help me and answer any and every question I ever came up with. I'd also like to thank Malin Planander, at Miljöbron, for invaluable input and support throughout the whole process.

Finally, I give my thanks to all my friends and family for cheering me on and giving the motivation I needed during this project. Especially Linnea, you are my Hero!

Leo Einarsson

Lund 2021

Contents

1 Introduction

This sections aims to explain the structure of the report, give background information and motivation for pursuing this project. It will also shed light on the purpose and aim set and any delimitations for achieving the objectives.

1.1 Outline of the report

The first section will describe the background and aims of this project. Next up will be the theory and method sections. These are separated into a district heating part - which factors affect heating and the typical trends, and a statistical part - explaining the how to apply the statistical bootstrap to regression and time series data and how to create confidence and prediction bands for these. The theory section will explain the basics of those concept, while the method section will delve deeper into how these are used in this project. In the method section, the model and data used will be explained as well. Finally, the results will be presented, followed by a discussion of the results in regards to the objectives.

1.2 Background

District heating is a means of providing heating of home and water for many individuals in a centralised way. Fredriksen and Werner writes in "Fjärrvärme: Teori, teknik och funktion" [1] that district heating (DH) has many advantages regarding economics and environmental impact. The authors explain that the idea of having heating centralised is an effective way of reducing the total energy production as there can be more larger production units, which usually are more efficient than smaller ones. However, since the facilities are responsible for many households, it is important that the production facilities produce the right amount of heat at the right time to lower the environmental impact.

There are two main ways of making the heat production more effective. One way is to design the parts of the energy producing cycle, i.e., an energy production plant, storage and all connecting pipes etc, in a more efficient way, c.f. [2]. The other way, which is the purpose of the Company that proposed this thesis, is to have better prediction of the heat demand for the upcoming days. For this study, the focus is going to be on the models used for forecasting heat production. Having a good forecast is essential for optimising heat production and minimising environmental impact. By having an accurate prediction, production companies will have a stronger basis when deciding which units to operate, in both near and distant future. Johansson writes in their doctoral dissertation about smart district heating, that it's usually expensive for production plants to start an extra unit, c.f. [3]. This is because all units have a static cost for starting heat production, in addition to a variable cost depending on the amount that is produced. Johansson also writes that the production plants usually have environmentally friendly and cheap heat sources as a base, but when the need is high, they would need to switch to more expensive sources, typically fossil fuel. Thus, it would be more economically and environmentally beneficial if one can

avoid using the expensive units as much as possible. To do that, the producers would need to know when they do not have to use the dirty fuel. Since district heating is the most common means of heating in Sweden, at around 50% of all heating according to Euroheat [4], lowering the emissions from the district heat production would likely help reduce the environmental impact.

A forecast for district heating could be considered to be good in different ways. Regardless, it would be beneficial to know within which boundaries the heat load is likely to stay within. This can be shown in the form of *confidence* and *prediction* intervals. Now, to make these intervals, one would usually gather data from many years, but by using the statistical bootstrapping technique, it is possible to generate artificial data with approximately the same properties as the original with a smaller sample set. Generating more bootstrap data then allows for statistical tests and creation of confidence and prediction intervals, which could give a better picture of the reliability for different prediction models. Bootstrapping is a powerful tool that can be used for analysing how good an estimate is. The classical bootstrap does require data to be independent, which data in a time series like heat load clearly is not. There are a few alterations or extensions to the classical bootstrap that does allow for the technique to be used for both regression analyses and time series data in which data is dependent in some ways.

The heat load varies quite a bit depending on the time of the year. As shown in Figure 1 below, the heat load is clearly higher during colder periods than the warmer ones. Joakim Henriksson and Sophie Rudén were tasked with improving a model for forecasting heat load for their master thesis and decided one way to do so was to look at the seasonal differences [5]. They concluded that they could create models with parameters better suited from each season, however, when applied to another season they were not as accurate. A possible conclusion could be that a model that changes its parameters throughout the year would be preferable.

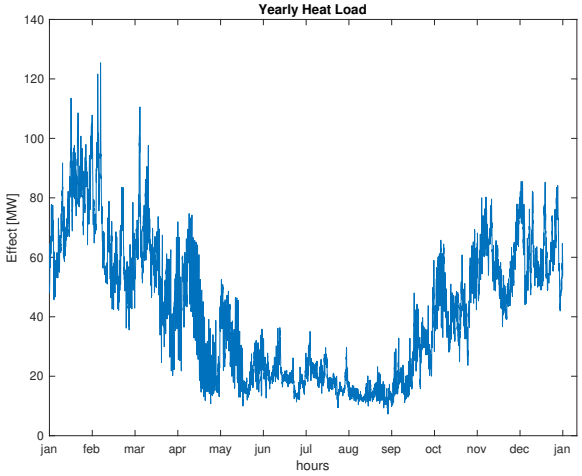


Figure 1: *The heat load at a heat production plant in Sweden during 2019*

A question then appears: could the forecasting be improved by having even narrower seasonal partitions? Henriksson and Rudén did find that their partitioning of four seasons performed better than just using the same model for the whole year [5]. Would it then become even more precise by going down to a specific model for each month, or even each week? This could possibly also depend on the time of the year. As seen in Figure 1, the period between June and September does not vary as much as April or October and perhaps those transitional periods between seasons could benefit more from smaller partitions?

1.3 Aim and Objectives

The aim is to gain a better understanding of how the partitions of the year and model parameters affect the reliability of the model. By using the current model of the company as a starting point for analysis, different partitions of the model parameters will be explored.

In order to arrive at an answer, the following objectives will be examined:

(i) Which partitions for the model are of interest?

(ii) Which partition gives the most reliable results for each season?

(iii) Which seasons are hardest to predict for the different partitions?

1.4 Delimitations

For this study, no proper distinction will be made between commercial and residential areas connected to the heat network. There will be more focus on the overall heat load for the heating network rather than individual housings.

There is a slight difference between produced and consumed heat as well, mostly caused by loss of heat when transferring the heat to the subscribers. For this project, however, only the heat production will be studied.

The model used by The Company have some real time updates of weather measurements and corrections done. In this project, only the base prediction will be studied and leave the add-ons out of the equation, as many of these functions are closely integrated in the company's software which I do not have access to make changes in. The weather data used, even for prediction, will be fully known as well, which is not the case when creating forecasts in reality, as the weather data also would carry additional variance.

Another limitation when it comes to the model, is that it cannot be fully disclosed. However, even if the exact model will not be shown here, its main properties and anything needed to understand the results will be explained.

2 Theory

In this section some theory regarding district heating and the statistical method of bootstrapping will be described. For district heating, its seasonal differences and known factors affecting the demand for it will be explained. The statistics part will explain the idea of bootstrapping, in general and for time series, and how to create confidence and prediction bands with it.

2.1 District Heating

Heat usage can be divided into two main usages. These are *heating of space* and *domestic hot water usage*. As its name suggest, the former alludes to the heat required to keep the temperature of a space within a comfortable level. This is usually the main source of heating a building, for instance radiators. The domestic hot water usage includes most things a person would need heating for other than the room temperature. These things could be hot water for a shower, washing machine or washing your hands. It is up to the heat producing plants to plan for and distribute the required heat for space heating and personal use.

There are two main factors which affect the heat load: weather and consumption pattern. The weather outside affects the need for space heating, while the consumption pattern provides mostly for hot water usage.

2.1.1 Weather factors

Weather influences the need for heat in many different ways. The strongest relation is with that of the outdoor temperature, while other factors such as wind, global influx and humidity are other, not quite as strongly connected. To better explain the seasonal and daily heat load trends, some of these factors shall be explained more thoroughly below.

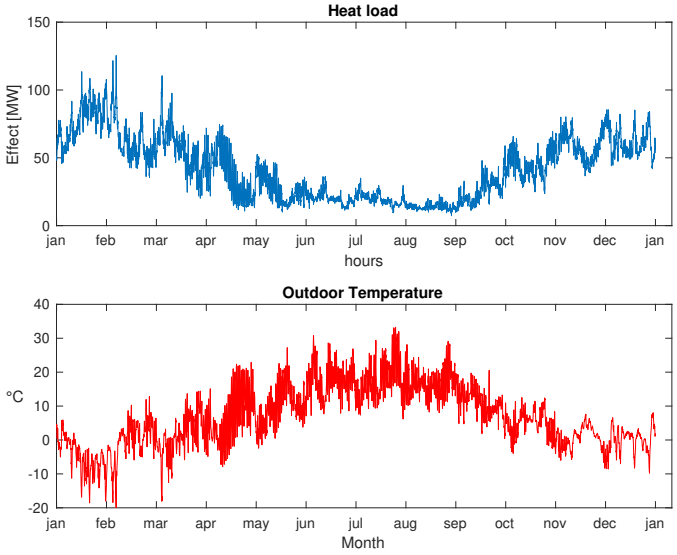


Figure 2: *Energy production and outdoor temperature measured in 2019 at a plant in Sweden, one point of data for every hour*

First we discuss the outdoor temperature. It has been shown that the heat consumption is closely tied to the temperature outside. From Figure 2, we see that the heat production is seemingly strongly correlated to the outside temperature most of the year. This is explained by Newton's law of cooling, which states that an object experiences heat loss linearly proportional to the difference of the object's and the surrounding's temperature. Mathematically it can be expressed as

$$Q = k \cdot A \cdot (T(t) - T_{out}), \tag{1}$$

where Q is the heat flow, A is the area of the object, $T(t)$ is the temperature of the object, in this case the indoor temperature of a building, at time t, T_{out} is the outdoor temperature and k is a heat transmission constant, also depending on the thickness of walls, insulation among other things in this case. Admittedly, when it comes to heating of buildings, this relation is not always linear. Frederiksen and Werner [1] writes that the walls have an inertness which allows for the walls to store some amount of heat. Because of this, the temperature from a few days ago can still affect the need for adding heat for the current day.

There is also a cut-off point for when the temperature affects the heat load. During at least half of the year, the outdoor temperature is lower than the desired one for housing. But there are a few months when heating definitely is not required and most people would prefer an air conditioner or a cooler. After a certain threshold in temperature outside, no space heating is required. Plotting a year's worth of hourly heat load measurements against the outdoor temperature, Figure 3 below shows how there is a clear cut-off temperature when the relation is no longer linear. Instead, after about 15°C the outdoor temperature doesn't affect the heat load nearly as much as before.

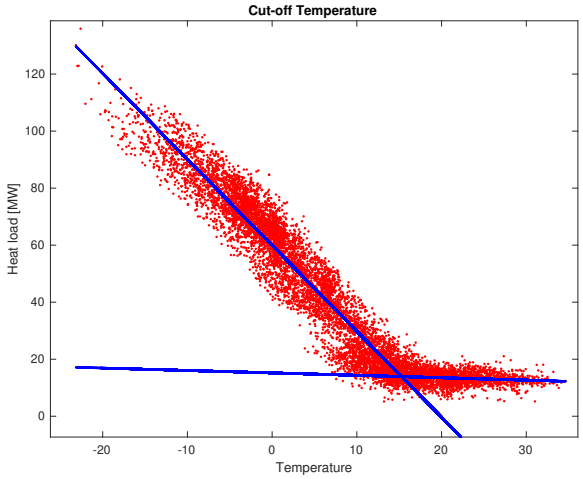


Figure 3: *Cut-off temperature for a typical heat producing plant. The produced heat load is plotted against the outdoor temperature. The cut-off temperature can be found at the intersection between the two lines.*

Wind speed can also affect the heat consumption. Normally ventilation is needed for exchanging stale air for fresh air, however, a greater wind speed can cause more drafts in unintentional places if isolated badly. This leads to more air from the outside to seep in, and if it is colder outside, even more heat would be required for keeping the indoors temperature comfortable, according to Frederiksen and Werner [1]. In the study conducted by Henriksson and Rudén [5] however, they found that the wind speed did not always relate significantly to the heat load. They did find that it had most effect during winter season, but could not draw any conclusion regarding its effect during spring, summer and autumn.

Finally, the global irradiance can also affect the heat demand. Global or solar irradiance refers to the added heat from direct sun light. Henriksson and Rudén [5] also examined the significance of global irradiance, during the studied summer period they found it to have a weak, negative effect on the accuracy of their model. Nothing could be said about the other seasons however.

There are of course other aspects that could affect the heat load as well. For instance humidity and wind direction, however, these will not be studied in this thesis.

2.1.2 Social Behaviours

Next we have consumption pattern, the social component which varies depending on the time of the day and the year. Just like the temperature changes depending on the time of the year, so does human behaviour. Not only does the sun heat up houses more in summer, but people also tend to use less heat for personal use, such as hot showers. This leads to an extra effect in general which makes the heat consumption even more differentiated between the colder and warmer months.

The personal usage creates a load pattern that differs throughout the day in accordance to social behaviours. In general, less heating is required at night and then there are two peaks, one in the morning, and one in the evening once people stop working. This effect is generally easier to spot during summer, as the space heating part is larger during colder temperature.

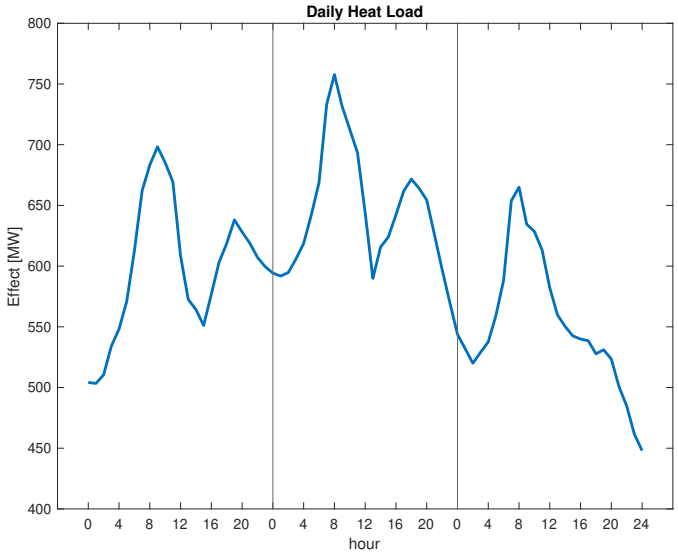


Figure 4: *Daily heat load for three days in January. There is a clear morning peak in heat load at around 7-9 o'clock. The rightmost day is a Saturday and the afternoon peak is practically not there at all*

There is also a pattern that varies depending on the day of the week. Two distinct categories that can be seen are those of weekdays and weekends. During weekdays, the majority of people get up between 6-8 o'clocko to prepare for work and take a shower or use hot water in other ways. In the same way, they return home around 16-18 to do washing or dishes etc. This leads to a pattern with a morning peak and an after work peak in heat usage. During weekends, these peaks are usually not as prominent as people distribute their hot water usage more spread out compared to the weekdays.

To summarise, there are a few different trends and auto-correlations to be found in district heating.

- **Daily (24h):** Morning peaks, not as much during night as most people are sleeping
- **Weekly (168h):** Weekends differ a bit from work days.
- **Yearly:** Seasonal dependence, warmer season = less heat required and produced.

2.2 Bootstrapping

2.2.1 Classic Bootstrapping

Ever heard of the paradox of pulling oneself up by ones bootstrap? This is it, but statistics.

For almost any statistical problem, there is always some uncertainties when applying statistical analysis on gathered data. For example, given a typical task of finding the mean height of all people in Sweden, the only way of getting the true distribution, F, and true mean, would be to receive information of the height of the entire population. That is usually not something that can be done without a great amount of resources and time. So what is usually done? - Getting a sample! Decide on a sample set of people, preferably as random as possible. Then perform a statistical analysis on the sample data and get an estimate, F_n, for the distribution and mean. That is where it could end. It may be a good estimate or it may be quite a bit off. Usually, for a good data set, the more data used, the closer to the true mean the estimate gets, but once again, it can be expensive to get that much data. And in the end, it's still just an estimate.

Enter Bootstrapping, a technique that can be used for determining how good an estimation is. J.S Urban Hjorth writes that a bootstrap often is used for measuring biases or uncertainties for estimated parameters [6]. They also mention that the bootstrap technique has developed quite a bit since and can nowadays be used for classification, regression, time series and other kinds of problems.

The original bootstrap method would require data points to be independent. Clearly, the data points in heat load are dependent and have properties of auto correlation, as it is a time series, however, the basic idea will be explained before going into more detail on how to work with bootstrap on more dependent data. The main idea of the traditional bootstrapping of individual data is to gather "more" information from ones sample by creating bootstrap data. Given a sample of n data points,

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_n), \quad x_i \in F,$$

where F is an unknown distribution, a first estimate $\hat{\theta}$ of the desired parameter θ can be created. This $\hat{\theta}$ can be used to describe an estimated distribution F_n which is known and from this more data can be "generated". It may sound preposterous, however, it has been proven to have statistical ground. The way this is done, is by *resampling*. Resampling is the process of creating alternate data sets by sampling the elements from the original data set, $\boldsymbol{x}$, with replacement and generating new bootstrapped data sets, $\boldsymbol{x}^*$. From each $\boldsymbol{x}^*$, a bootstrap parameter, θ^*, can be computed the same way $\hat{\theta}$ was estimated from $\boldsymbol{x}$. A visualisation of resampling with four data points is shown in Figure 5 below.

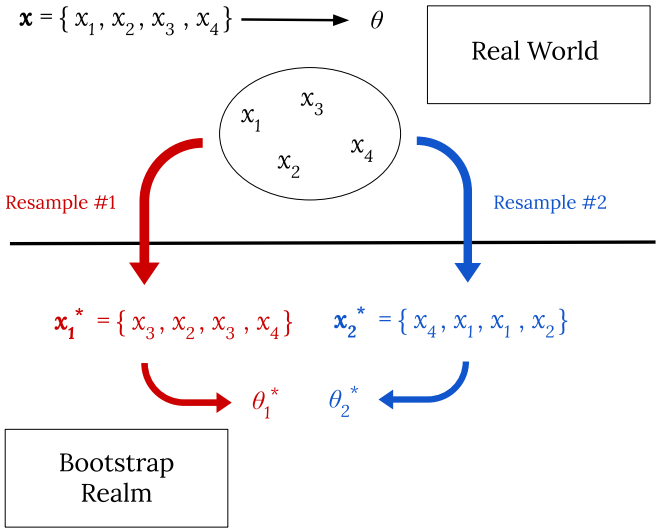


Figure 5: *Illustrating how bootstrap samples* $\boldsymbol{x_i^*}$ *and parameters* $\boldsymbol{\theta_i^*}$ *are created by resampling*

What is even more amazing, is the fact that by gaining more information about $\hat{\theta}$, more information can be gained about the true parameter. This is because the relation between θ and $\hat{\theta}$ is approximately the same as the one between θ^* and $\hat{\theta}$ when a sufficient amount of bootstrap estimates have been created. The relation looks as follows:

$$\hat{\theta} - \theta \approx \theta^* - \hat{\theta}. \tag{2}$$

This approximation gets closer, the more bootstrap samples that have been created. This report will not delve further in why this is correct, you may read about it on your own.

2.2.2 Bootstrap for regression - Residual sampling

When it comes to regression, the measurements alone are not what becomes bootstrapped. Since the goal is to find a pattern between input and output data, it would yield chaotic results if the measurements were scrambled at random. Instead, there are two options: bootstrap pairs of (x_i, y_i) or perform bootstrap on the residuals, ε_i [6]. For this study, residual resampling is going to be performed.

From here on, a bold symbol indicates a vector or a matrix. Assume we have n points of data and p number of parameters. If using least squared estimation for estimating the parameters, $\boldsymbol{\theta}$ of size $p \times 1$, the estimate, $\hat{\boldsymbol{\theta}}$, is given by

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{y}, \tag{3}$$

in which $\boldsymbol{x}$ is a matrix of size $n \times p$ consisting of input data corresponding to the parameters, for instance current temperature and social components, and $\boldsymbol{y}$ is a

vector of output data, the produced heat load, of size $n \times 1$. This in turn means we can express $\boldsymbol{y}$ as

$$\boldsymbol{y} = \boldsymbol{x}\hat{\boldsymbol{\theta}} + \boldsymbol{\varepsilon}, \tag{4}$$

where $\boldsymbol{\varepsilon}$ are the residuals.

Using the first estimation of the parameters, $\hat{\boldsymbol{\theta}}$, and computing an estimated dimension space for the data, $\hat{\boldsymbol{y}}$, sample the residuals, $\boldsymbol{\varepsilon}$ between the true data and estimation can be extracted. Then the residuals can be bootstrapped

$$\varepsilon_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n \tag{5}$$

The bootstrapping of the residuals is done the same way as any other, pick out a random residual, with replacement, for each data point, giving you $\varepsilon*$. Then bootstrap data, $\tilde{y}_i$ can be created by adding the residuals to the estimated data.

$$\tilde{y}_i = \hat{y}_i + \varepsilon_i^* \tag{6}$$

After that, bootstrapped parameters, θ^*, can be estimated the same way $\hat{\theta}$ was, however, $\tilde{y}$ is used instead of y.

$$\boldsymbol{\theta^*} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\tilde{\boldsymbol{y}} \tag{7}$$

From the new $\boldsymbol{\theta^*}$, new bootstrap estimations, $\boldsymbol{y^*}$ can be created from

$$\boldsymbol{y^*} = \boldsymbol{x^T\theta^*}. \tag{8}$$

By doing this many many times, a good approximation of the true parameters $\boldsymbol{\theta}$ can be found as through relation (2).

As a side note, most literature denote what I call $\tilde{y}$ as y^*. However, in this work, some analysis is going to be preformed using the resulting heat load y^* as I describe it, which is why the notations may differ from others' works.

2.2.3 Bootstrapping for Time Series

A time series, like the heat load, does not consist of independent data points. There is auto-correlation within different time frames and when the data is sampled is an important factor. This means resampling the residuals completely at random may not replicate the time series aspect of the data particularly well.

It's still possible to reuse much of the theory for regression by applying *Block Bootstrapping*. Fanny Bergström wrote in their bachelor thesis about how to apply this when using bootstrap on time series [7]. The difference, as the name suggests, is that data is divided into blocks instead of being seen as isolated points. The idea is to preserve some of the dependency withing the data set by resampling in groups. Bergström describes a number of different ways to create these blocks:

non-overlapping (NBB), moving (MBB), circular (CBB) and stationary (SB) block-bootstrapping. They also found that while there's still dependency loss regardless of method used, MBB and CBB generally led to a smaller error, when applied to a time series. In short, by dividing the data and computing the residuals $\boldsymbol{\varepsilon}$ in blocks, some dependency can be retained.

Both MBB and CBB uses overlapping blocks. Both methods divides the data into blocks of consistent size, though they may overlap. Assuming there are n points of data, a block length l in chosen such that $l|n$ for easier concatenation of blocks. Then $n - l + 1$ blocks of size l are created, where each of them is shifted one step at a time.

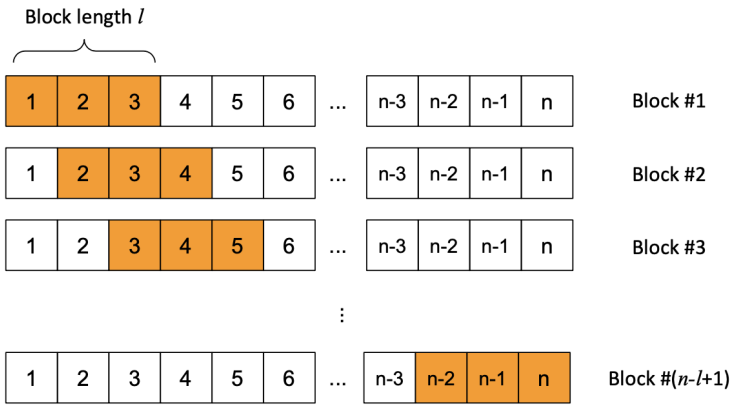


Figure 6: *Moving Block Bootstrap with n data points and a block length of $l = 3$.*

One downside to creating the blocks through MBB is that the samples near the edges won't be present in the same amount of blocks as those closer the middle. Bergström explains that CBB circumvents this by connecting the first and last block to form a cycle [7]. It is useful for when the data points loop back on their own in a sense, for instance yearly temperature. However, each month or season does not loop back into themselves, which means the residuals could be misleading.

2.2.4 Confidence and Prediction Bands

When studying how accurate the model within each parameter partition, a confidence or prediction band can be utilised. Since we are dealing with regression, we will use confidence bands as opposed to pointwise confidence intervals. A confidence band consists of an upper and lower limit for in which the whole regression line is going to stay between for a certain percentage of the time. To create this band, we need to determine an upper and lower limit, $f(x)_{upper}$ and $f(x)_{lower}$ respectively, such that

$$P(f(x)_{lower} < f(x) < f(x)_{upper}, \forall\, x) = 1 - \alpha$$

In this case, $f(x)$ represents the heat load y and $1 - \alpha$ is the confidence level. To determine the bounds, a constant, d, can be computed and be added or subtracted

to $\hat{y}$,

$$d_i^* = \sup_i |y_i^* - \hat{y}_i| \tag{9}$$

in which $i = 1, \ldots, n$ and represents the heat load at hour i and n is the number of hours in total. When doing these for bootstrapped data, the confidence band comes from computing (9) for all N bootstrap estimations

$$\begin{cases} d_1^* & = \sup_i |y_i^{*(1)} - \hat{y}_i| \\ & \vdots \\ d_N^* & = \sup_i |y_i^{*(N)} - \hat{y}_i| \end{cases} \tag{10}$$

and then create a histogram of the supremum error for all bootstrap estimations d. From the histogram, we can find d_α as the value in the histogram such that it is larger than $(1 - \alpha)$ percent of all the d-values.

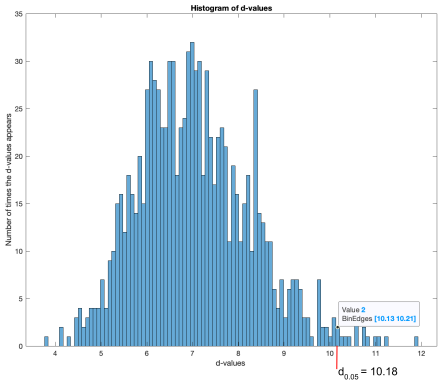


Figure 7: *Histogram of 1000 d-values with a confidence level* $(1 - \alpha) = 0.95$

The upper and lower limit are then created as the estimation with d_α added and subtracted respectively as follows:

$$[\hat{y} - d_\alpha, \hat{y} + d_\alpha]. \tag{11}$$

The value of the true heat load y is then likely to be within these boundaries $(1 - \alpha)$ percent of the time. Since every partition is using their own parameter set and model, we will need to compute d_α for each partition. We can then use these values to visualise the confidence band and give an estimation of how spread the district heating is for each partition. To summarise the variance for each partition, we will divide d_α with the mean of the heat load for each partition, to get an idea of how varied they are.

Utilising confidence bands gives a picture on how robust the model is, however, for future values, we need prediction intervals or bands. The process of creating

prediction intervals and bands are similar to those of confidence but it uses a different error as residuals. Stine explains the process of creating prediction intervals for regression using prediction errors [8].

Let $\boldsymbol{x_f}$ be a matrix of future input data. Then the future prediction $\hat{\boldsymbol{y}}_{\boldsymbol{f}}$ can be computed as

$$\hat{\boldsymbol{y}}_{\boldsymbol{f}} = \boldsymbol{x}_{\boldsymbol{f}}^{\boldsymbol{T}}\hat{\boldsymbol{\theta}}, \tag{12}$$

while the true future heat load, $\boldsymbol{y}_f$ would be

$$\boldsymbol{y_f} = \boldsymbol{x}_{\boldsymbol{f}}^{\boldsymbol{T}}\boldsymbol{\theta} + \boldsymbol{\varepsilon_f}, \tag{13}$$

where $\boldsymbol{\theta}$ is the true parameter values, yet unknown to us, and $\boldsymbol{\varepsilon_f}$ is the future residuals.

We want to construct a prediction interval such that the estimated value of y_{f_i} is contained within the interval with a probability of β. Mathematically it can be expressed as the following

$$P_{y_f}(y_f \in I(y_f)) = E_y(1\{y_f \in I(y_f)\}) = \beta, \tag{14}$$

where $1\{y_f \in I(y_f)$ is the probability that this is true.

Now, to construct an interval for the bootstrap, we are going to use the estimate $\hat{\boldsymbol{y}}_{\boldsymbol{f}}$ and add an upper and lower bound to it. Thus the interval would be created as

$$I^B(y_f) = (\hat{y}_f) + l_f^B, \hat{y}_f + u_f^B, \tag{15}$$

where l_f^B and u_f^B respectively are the lower and upper bonds chosen such that the expected value that the bootstrapped heat load

$$\tilde{y}_{i,f} = x_{i,f}\hat{\boldsymbol{\theta}} + \varepsilon_{i,f},$$

is contained within the bounds with a probability β. Mathematically expressed as

$$E_*(1\{\tilde{y}_f \in I^B(\tilde{y}_f^*)\}). \tag{16}$$

To decide l_f^B and u_f^B, we need to take a look at the prediction error, rather than the bootstrap residuals as we do for the confidence bands. In other words, instead of studying ε^*, we use

$$\boldsymbol{D}_{\boldsymbol{f}}^{\boldsymbol{*(n)}} = \tilde{\boldsymbol{y}}_{\boldsymbol{f}}^{\boldsymbol{(n)}} - \boldsymbol{y}_{\boldsymbol{f}}^{\boldsymbol{*(n)}}, \tag{17}$$

in which $\boldsymbol{y}_{\boldsymbol{f}}^{\boldsymbol{*}} = \boldsymbol{x}_{\boldsymbol{f}}^{\boldsymbol{T}}\boldsymbol{\theta}^{\boldsymbol{*}}$, (n) indicates the $n:th$ bootstrap sample and $\boldsymbol{D}$ is a $N \times T$ matrix. By computing $\boldsymbol{D}_{\boldsymbol{f}}^{\boldsymbol{*}}$ for every bootstrap sample, we get

$$\boldsymbol{D}_{\boldsymbol{f}}^{\boldsymbol{*(1)}}, ..., \boldsymbol{D}_{\boldsymbol{f}}^{\boldsymbol{*(1)}}, \tag{18}$$

which we can use to analyse the distribution for every point in time, t, and create a prediction interval for those by explicitly finding the values for which $1-\beta$ percent of the values are contained within. Either we can keep them as pointwise intervals or we can compute a prediction band for each partition, by the same principle as for confidence bands, see equation (10).

3 Method

In this section, the data used and choices made for bootstrapping and analysis will be explained. First is a bit of data chosen and a description of how it was preprocessed. Next will be a short overview of the model and then which parameter sets that were chosen to be studied. Finally, the evaluation method will be described, along with how the bootstrap was performed.

3.1 Data used in study

There has also been some amount of preprocessing made to both heat load and weather data. Sometimes the measuring instruments for both heat load and weather data get readings that are missing entirely or deemed to be outliers. For those points, the data was adjusted or added manually. When it comes to temperature, as shown in Figure 3, the temperature is not as important for the demand of heat load after a certain threshold. As such, temperature values over that cut-off point are reduced to a lower temperature.

The data used in this project has been collected by The Company and consists of three customers. One of these are located in Sweden (S), one in Denmark (D) and the last one in Switzerland (C). For all of these, hourly measurements of external input and heat load production will be used. For these plants, data from two years will be used, one for estimating the parameters and confidence analysis, and one for prediction and prediction analysis.

Plant S only use outdoor temperature as external input and we have used data points from two years, 2018 and 2019.

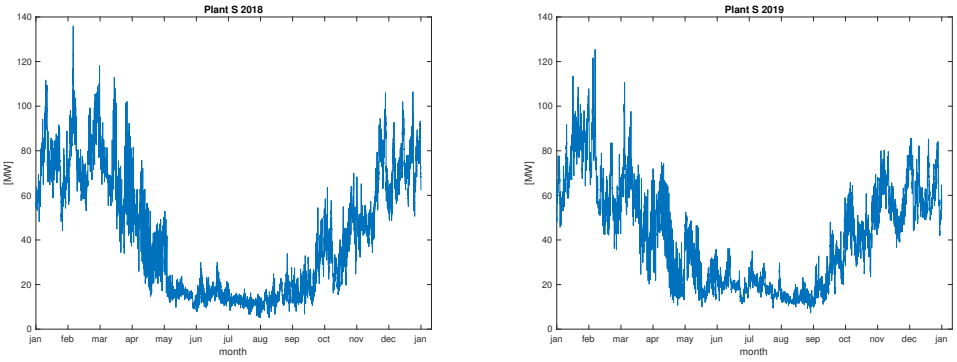


Figure 8: *Heat load at plant S during 2018 to the left and 2019 to the right.*

Plant D use wind speed and global influx in addition to the outdoor temperature as input. The data points used are from two years, 2019 and 2020.

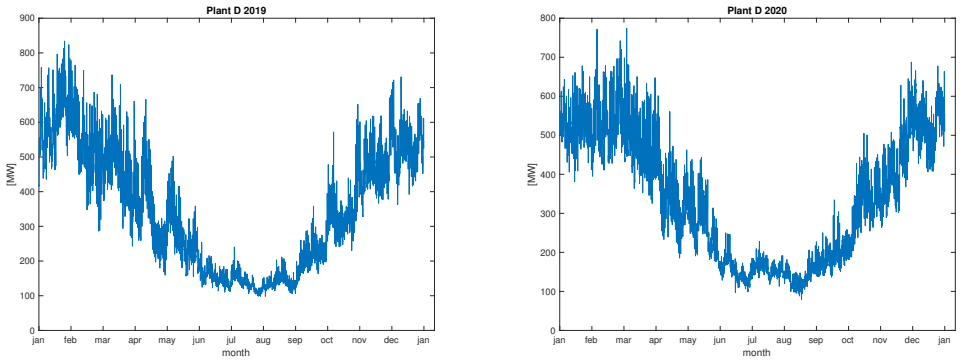


Figure 9: *Heat load at plant D during 2019 to the left, and 2020 to the right.*

Plant C only uses outdoor temperature as input. We have access to data points from the same two years as for plant D, e.g. 2019 and 2020.

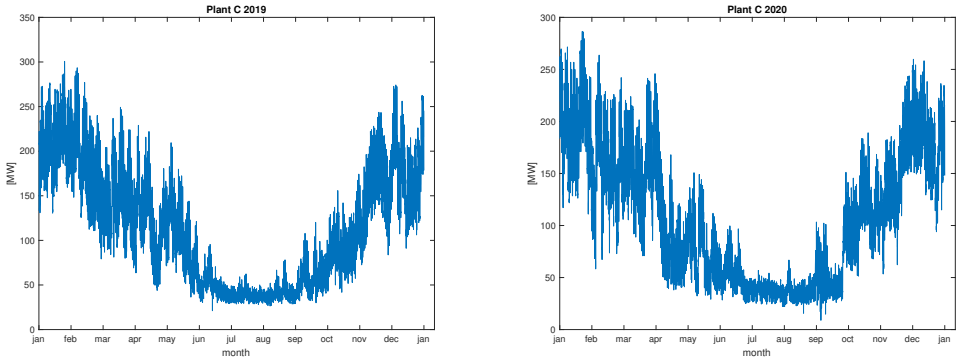


Figure 10: *Heat load at plant C during 2019 to the left, and 2020 to the right.*

To get a better overview of how the heat load varied from the year the model was fitted to and the year that is to be predicted, Figure 11 below shows both preprocessed heat load and temperature for both years plotted together. These can also be found in greater resolution in Appendix A.

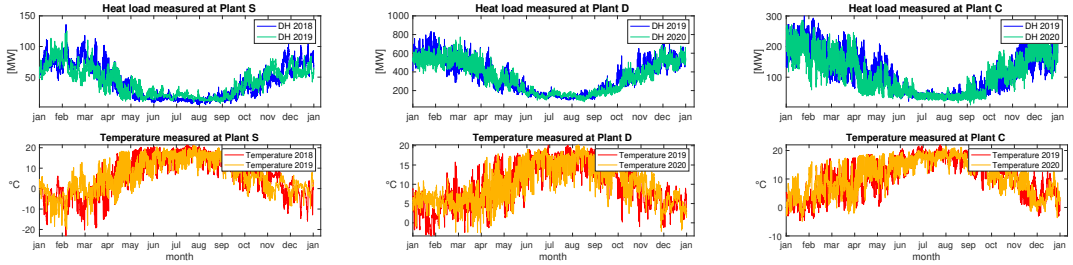


Figure 11: *This depicts the district heat load and temperature for both the year used for fitting the model and for predicting. The graphs to the left illustrates the data for plant S, the middle for plant D and the right one for plant C. The dark blue graphs are the heat load the model is fitted to, while the green graph is the heat load we try to predict. The deep red lines are last year's temperature and the bright orange is for the predicted year.*

3.2 Overview of Model

The model is a combination of the two separate factors related to district heating. One part of the model regards the weather data, i.e. outdoor temperature, optionally adding wind and solar influx, as input. The other part models social behaviour depending on the hour of the week. Linearly combining these two parts gives the estimated heat load $Q(t)$ for the time t,

$$Q(t) = \boldsymbol{\theta_1} \cdot Input_{weather}(t) + \boldsymbol{\theta_2} \cdot Input_{social}(t) \tag{19}$$

in which $Input_{weather}(t)$ and $Input_{social}(t)$ contain terms that are multiplied with different coefficients (parameters) $\boldsymbol{\theta_i} = \{\theta_{i,1}, \theta_{i,2} \dots\}$. It is basically a combination of two linear models with different input data. These parameters are recomputed for each partition of the year to give the model better characteristics for the period they are used for. This is done by fitting the model based using least squared estimation, as seen in Section 2.2.2.

3.3 Parameter Sets to be analysed

Henriksson and Rudén discovered that the studied model seemed to benefit from being more specialised depending on the current season [5]. They tested their model, using different parameters for four consecutive weeks in January, April, July and October, but by going for even greater resolution and have weekly or monthly parameter sets instead, would the accuracy become even greater? The more narrow partitioning could hypothetically help with the "Transition Seasons", the periods when winter changes into spring and so on. During these periods, the temperature can change quite a lot, which a model with few partitions is not always good at keeping up with. Just think of a classic April month, usually the weather and temperature change drastically from the first to the last day of the month.

Just by trying out some different partitioning sets, visually getting a first impression of how the intervals affected the model, some particular sets to investigate was decided and follows in the table below:

Table 1: *A table of all different parameter partitions to be analysed*

Number of Parameter partitions	Days before parameter switching	Description
1	[365]	One parameter set for the whole year.
4	$[90, 91, 92, 92]$	One parameter set per Season, starting in Jan-Mar.
12	$[31, 28, 31, 30, 31, 30, 31, 31, 30, 31, 30, 31]$	One parameter set per Month
52	$[8, 7, 7, 7, \ldots, 7]$	8 days for first and 7 for the rest.

If there is a leap year, one day of data is removed. That day was chosen to a day in July that was missing measured data for plant D, and the 31:st of December for plant C. No data from plant S used occurred during a leap year, so no days were removed.

3.4 Evaluation method

To analyse the reliability of the different partition models, the statistical bootstrap will be applied. Once the bootstrap data is created, it will then be used to construct confidence and prediction bands. With prediction, some pointwise intervals may also be constructed to examine morning peaks in more detail.

To evaluate the predictions, we will be using root mean squared percentage error (RMSPE). This is to get a better grasp at how large the error of the prediction is in relation to the amount of heat load produced. The RMSPE is computed as

$$\sqrt{\frac{1}{n}\sum_{i=0}^{n-1}\left(\frac{y_{i,f}-\hat{y}_{i,f}}{y_{i,f}}\right)^2}\cdot 100, \tag{20}$$

in which $y_{i,f}$ is the heat load for the predicted year at hour i and $\hat{y}_{i,f}$ the estimated heat load from the model for the same point in time. This will be applied both for the whole year, but also for shorter intervals.

When studying the confidence and prediction bands, a similar approach will be used. By computing a value from the size of either band and divide it by the mean of the

heat load produced for the corresponding interval, a relative size of the bands can be gained and compared with bands for other intervals through the year.

The confidence and prediction levels chosen for the bands are 95% and 90% respectively.

3.4.1 Bootstrapping

Since most of the partitions are periods that doesn't necessary have the exact same properties at the start and the end, the non-circular MBB will be used. When it comes to block size, it's not very easy to pick one. As mentioned in section 2, daily variations are quite common in heat load and by picking a block length that covers 24 hours, some of that dependency should be preserved. It is worth to note that this will probably mean a loss of information regarding weekdays and weekends, but in order to combine the residual blocks, its advantageous if the block size is a divisor to the number of hours in a partition. Which is always the case with a day (24h), but not with a week (168h).

The Bootstrap Algorithm will have to be done separately for each partition, as they use different parameters and looks as follows:

1. Compute the estimate of the parameters using LSE, $\hat{\boldsymbol{\theta}} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{y}$.
2. Compute the estimated heat load $\hat{\boldsymbol{y}}$ using $\hat{\boldsymbol{\theta}}$ according to the model, $\hat{\boldsymbol{y}} = \boldsymbol{x}\hat{\boldsymbol{\theta}}$
3. Extract the residuals of every data point $\varepsilon_i = y_i - \hat{y}_i$
4. Create Blocks of residuals of length l.
5. Create a bootstrapped heat load, $\tilde{y}_i = \hat{y}_i + \varepsilon_i$, by resampling the residual blocks, concatenate these and add to $\hat{y}$.
6. Compute new model parameters θ^* by same method as for $\hat{\theta}$ but trying to fit to $\tilde{y}$: $\boldsymbol{\theta^*} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\tilde{\boldsymbol{y}}$.
7. Compute bootstrapped heat load y^* by using θ^* in the model: $\boldsymbol{y^*} = \boldsymbol{x}\boldsymbol{\theta^*}$
8. Repeat step 5-7 N times

Once repeated, there will be N bootstrap samples that can be used for analysis via testing and confidence and prediction bands. Using the bootstrap relation (2), by creating histograms and computing $d^* = y_i^* - \hat{y}_i$ and $\boldsymbol{D} = \tilde{\boldsymbol{y}}_{\boldsymbol{f}}^{(\boldsymbol{n})} - \boldsymbol{y}_{\boldsymbol{f}}^{*(\boldsymbol{n})}$ to create confidence and prediction bands respectively.

For this project, we have chosen to go with $N = 1000$.

4 Results

Here follows the result from the different partitions and estimation methods. First are the results from the confidence bands and fitting of the model to current year. Next will be the results regarding prediction and prediction bands. For some of the graphs in greater resolution, refer to the Appendix section.

4.1 Model fitting and prediction

In this section, the RMSPE will first be shown for the whole year for all plants both for fitting and prediction. Next RMSPE will be shown separately for each plant on a monthly basis.

Table 2: *This table depicts the RMSPE between y and $\hat{y}$ for the whole year. Note that the more parameter sets used throughout the year, the closer the model is able to fit to the true heat load.*

Parameter set	Plant S	Plant D	Plant C
Yearly	23.6152	18.3110	25.2945
Seasonal	16.9044	8.4952	12.9518
Monthly	12.3983	6.0301	8.5598
Weekly	8.8530	4.3346	6.8030

Table 3: *This table depicts the RMSPE between y_f and $\hat{y}_f$ computed for the whole year for all plants and all parameter sets tried.*

Parameter set	Plant S	Plant D	Plant C
Yearly	22.1592	20.0147	26.2690
Seasonal	17.7356	10.4363	15.5252
Monthly	16.6266	8.5739	14.4330
Weekly	32.0808	14.6482	22.3950

In Table 3 above, we note that the yearly and weekly parameter partitioning have the lowest accuracy when looking at RMSPE. On the other hand, the monthly parameter set have the least error in this regard for all plants. Additionally, plant D seems to have the lowest relative error of the tested plants.

In the following tables, Table 4 - 6, the RMSPE for all parameter sets will be shown on a monthly basis. Each table shows the results for the plants individually. The green and red values mark the lowest and highest RMSPE respectively for each parameter set, while the bold text represents which set had the lowest RMSPE for that month.

Table 4: *This displays the RMSPE on monthly intervals for plant S for all different parameter sets tested in the study. No specific parameter set had the lowest score for all periods, however, the monthly set performed best for the middle of the year, while the seasonal predicted more accurately in November and December. The yearly and weekly version were worse for predicting during summer, but the yearly did better during the winter months.*

	Yearly	**Season**	**Monthly**	**Weekly**
Jan	5.0568	**4.8413**	5.1559	12.4005
Feb	6.3373	**6.1795**	6.3861	12.7895
Mar	9.9597	9.2222	**7.6153**	23.0580
Apr	19.4618	**15.5138**	15.5722	25.0310
Maj	32.0300	24.2810	**23.5494**	40.1231
Jun	38.3404	33.9816	**27.8822**	41.5085
Jul	37.8839	**27.4392**	28.2402	58.5034
Aug	26.2504	20.0405	**14.5191**	34.2045
Sep	22.5914	14.9536	**13.8078**	40.8262
Oct	11.3570	12.6045	**10.2631**	13.1541
Nov	**6.2072**	7.8355	12.7393	20.0930
Dec	**4.5757**	5.6052	11.6914	25.1324

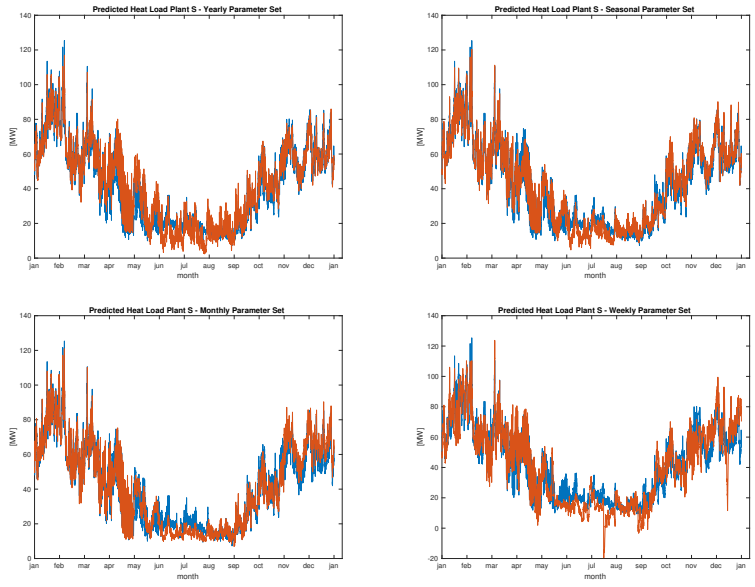


Figure 12: *Prediction and true DH for plant S. Upper left is yearly parameter set, upper right is seasonal, lower left is monthly and finally lower right is weekly.*

To summarise the predictions for plant S, all of the estimations seem to have a lower accuracy for May-July, but better results during winter. The yearly prediction in Figure 12 does appear to be a bit low during the first part of the year and too oscillating during summer. The weekly prediction seemingly have some odd peaks and dips throughout the year.

Table 5: *This shows the RMSPE on monthly intervals for plant D for all different parameter sets tested in the study. The monthly set performed the best in this regard for almost all months, although the seasonal set did predict slightly better in May and September. The weekly set predicted more accurately in July than June and August.*

	Yearly	**Season**	**Monthly**	**Weekly**
Jan	7.7988	5.1775	**5.0261**	14.3579
Feb	8.0512	**7.6448**	8.0884	10.4875
Mar	6.9024	**5.6604**	6.0895	10.0616
Apr	10.1245	**7.5936**	8.3189	24.8056
Maj	14.4020	**9.0137**	10.4418	15.2228
Jun	40.0546	22.3983	**12.1016**	16.8112
Jul	21.4270	8.5042	**7.0578**	9.5872
Aug	40.1294	15.0456	**10.6259**	17.5737
Sep	22.2210	**8.0257**	8.2709	15.1002
Oct	9.3077	11.0318	**8.2123**	13.7483
Nov	**5.9873**	6.8844	6.7804	10.5369
Dec	4.6571	4.4293	**3.8311**	10.0542

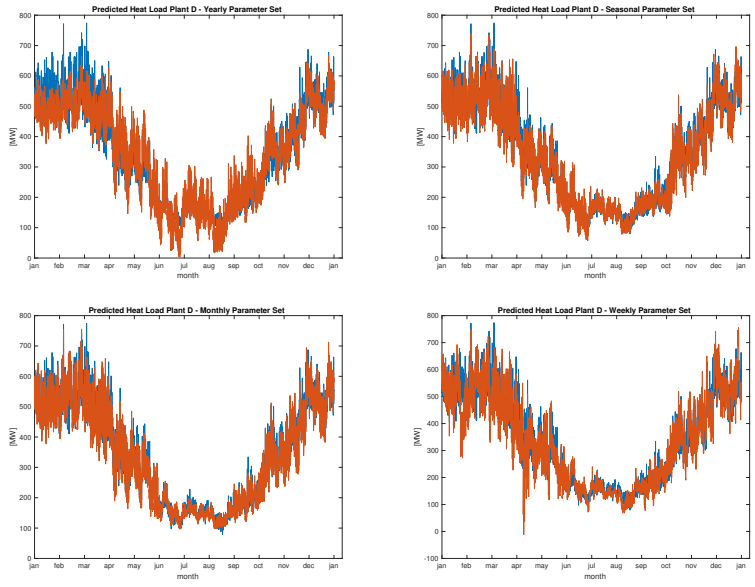


Figure 13: *Prediction and true DH for plant D. Upper left is yearly parameter set, upper right is seasonal, lower left is monthly and finally lower right is weekly. Overall the predictions with seasonal and monthly parameter set seem to be predicting closer overall, while the yearly seem a bit low in the first months and and a bit too oscillating in summer.*

In Figure 13 we see that the prediction with yearly parameters had trouble both at the beginning of the year and during summer. It seems all of them have more differing estimations during summer, especially June and August, though the model with a monthly parameter set seems to be the most accurate one.

Table 6: *This shows the RMSPE on monthly intervals for plant C for all different parameter sets tested in the study. Just as for plant D, the seasonal set performed better during May and September, whereas the monthly set generally predicted better at other times. Even here the yearly prediction is least accurate in the summer, but performs admirably during winter.*

	Yearly	Seasonal	Monthly	Weekly
Jan	8.0618	**6.5662**	7.2396	8.4893
Feb	**10.3516**	11.1265	11.7778	12.3991
Mar	8.7442	9.5823	**7.8686**	8.9748
Apr	21.1684	**18.0928**	18.6043	33.4647
Maj	25.0202	**19.8049**	22.0587	28.9231
Jun	30.0194	24.2276	**15.5990**	16.4639
Jul	43.0137	12.5992	**10.0933**	11.9356
Aug	48.8849	15.6765	**13.5382**	23.6896
Sep	39.4965	25.8835	**25.8510**	48.4595
Oct	9.9670	12.0314	**9.8816**	15.8143
Nov	**7.7889**	8.9522	9.8086	13.0413
Dec	6.8780	6.2432	**5.7824**	6.3372

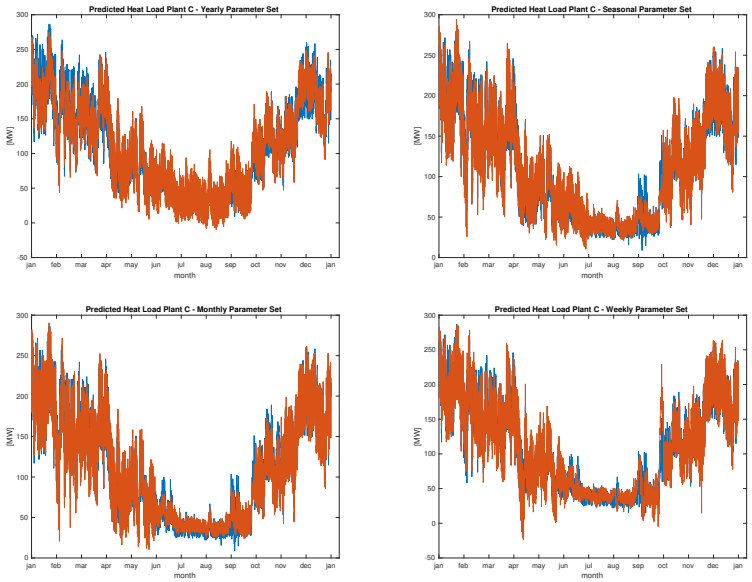


Figure 14: *Prediction and true DH for plant C. Upper left is yearly parameter set, upper right is seasonal, lower left is monthly and finally lower right is weekly. There seem to be some spikes in heat load in September that none of the predictions hit, except possibly the yearly model.*

As with Plant D, the estimation with a yearly parameter set for Plant C differs more than the others at the beginning of the year and the summer, as can be seen in Figure 14. September have the largest RMSPE for the seasonal, monthly and weekly model, and quite large for the yearly version as well.

4.2 Confidence and prediction bands

To show the results regarding the created confidence bands, the absolute values are not particularly telling. Since the plants are of different sizes and locations, the magnitude of the produced heat load also differs. To get a better overview of the relation between the heat load and confidence bands, let d_α be the magnitude of half the confidence band as in equation (11) and let $\bar{y}$ be the mean of the heat load. Then the numbers presented will be

$$\frac{d_\alpha}{\bar{y}},$$

which is how large the confidence band is in proportion to the mean heat load. This is done for every partition. For all tables in this section, except for the one with yearly parameter set, the relatively smallest confidence and prediction band will be coloured in green and the largest in red. Do note that the size of both the confidence and prediction bands have been divided with the mean of the predicted heat load for respective period and plant.

All models will have additional graphs depicting the prediction and both bands for an arbitrary week in June, September and December. For all figures illustrating the bands, the following legend will apply:

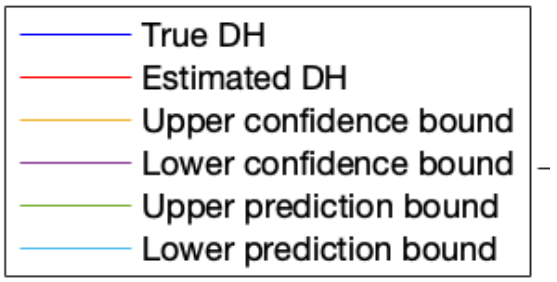


Figure 15: *The legend for all following graphs depicting the confidence and prediction bands.*

4.2.1 Yearly set

Here follows the results of the yearly predictions with and without confidence and prediction bands. Since the bands are of same size, they are relatively large during summer and oscillates more than the actual heat load which can be seen in Figure 16 and 17.

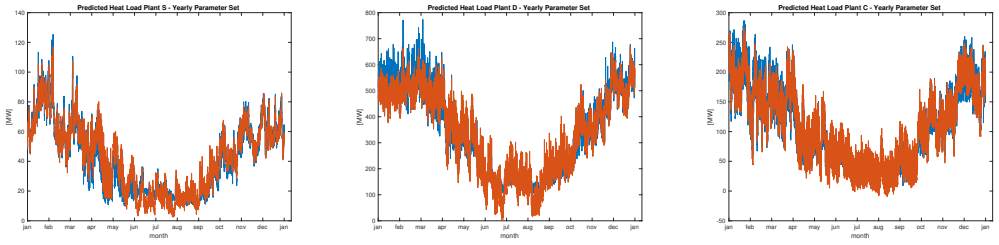


Figure 16: *Prediction and true heat load for all plants using yearly parameters*

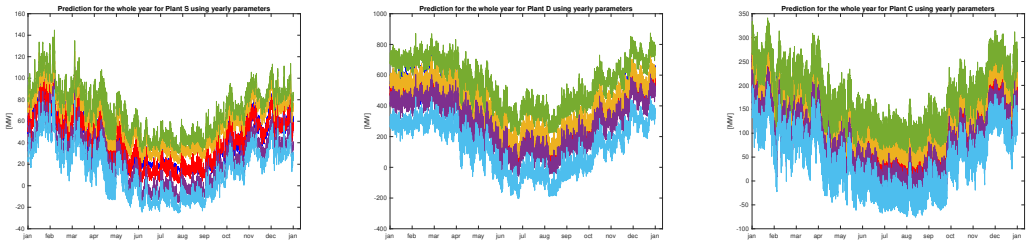


Figure 17: *Prediction and confidence bands added to the estimation and true DH for yearly parameter for all plants*

Table 7 shows the size of the bands relative to the heat produced. The confidence bands are clearly narrower for plant D than the other two, however, the prediction bands are of the same relative size.

Table 7: *This table shows the size of the 95% confidence bands and 90% prediction band in relation to the magnitude of the heat load for a single yearly set of parameters.*

(a) *Confidence 95%*

Period \Plant	**S**	**D**	**C**
Whole year	0.4293	0.1801	0.3346

(b) *Prediction 90%*

S	**D**	**C**
0.6542	0.6234	0.6343

In Figure 18, 19 and 20 we can see the head load and corresponding bands for a week in June, September and December for all plants. The bands are of the same size all year, however, the prediction appears to predict the shape of true heat load closer in December than June. September has some mixed results while the prediction oscillates too much in June.

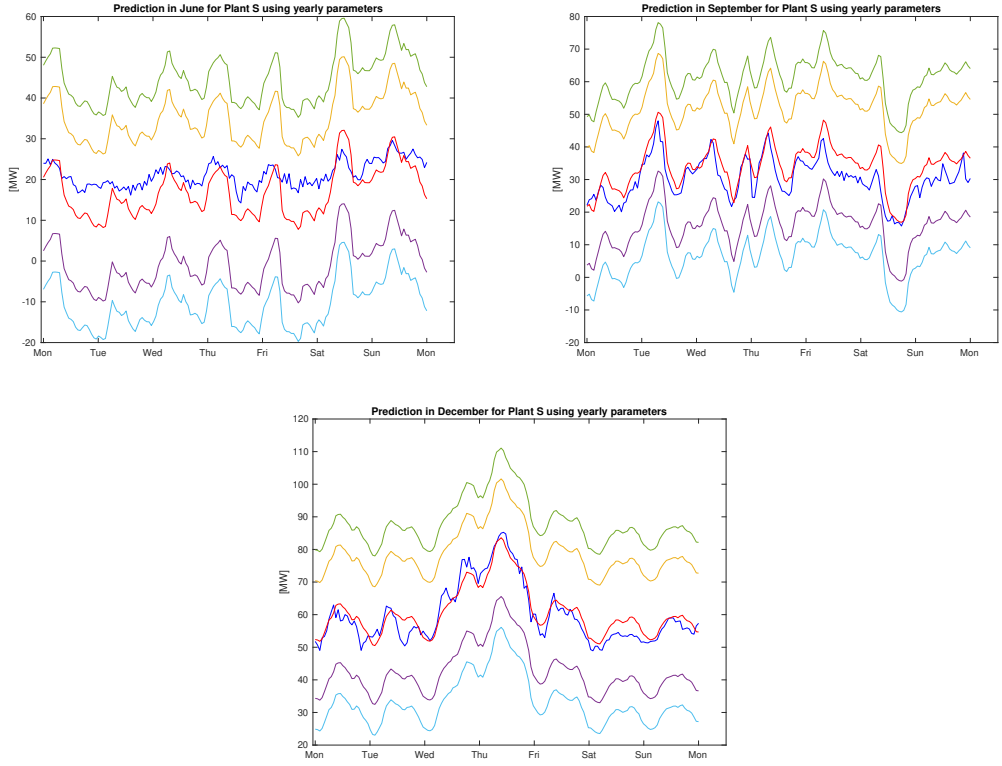


Figure 18: *Predicted heat load with confidence and prediction bands for plant S using yearly parameters. The periods shown are a week in June, September and December. The prediction in June is slightly low and oscillates more than the true heat load, the lower band is also almost constantly below 0 MW. The prediction is closer in both September and December, yet the lower prediction bound is still below 0 MW for a few points in September.*

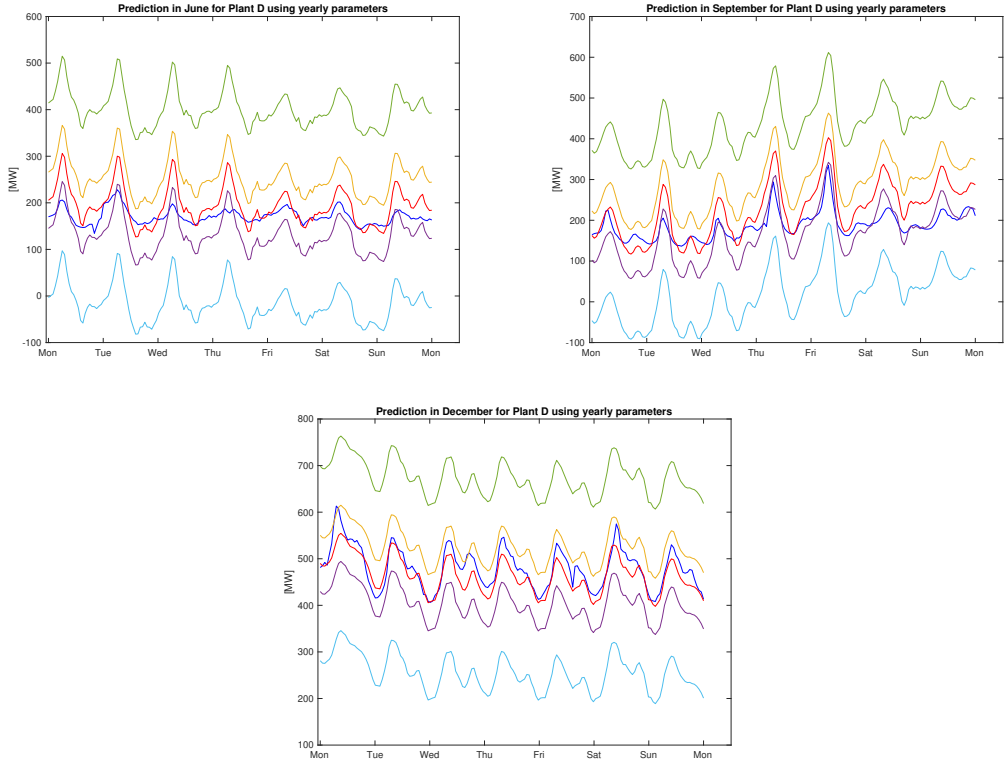


Figure 19: *Predicted heat load with confidence and prediction bands for plant D using yearly parameters. The periods shown are a week in June, September and December. The prediction in June is oscillating stronger than the true heat load and a over shoots a bit in September. The confidence band is much narrower than the prediction band.*

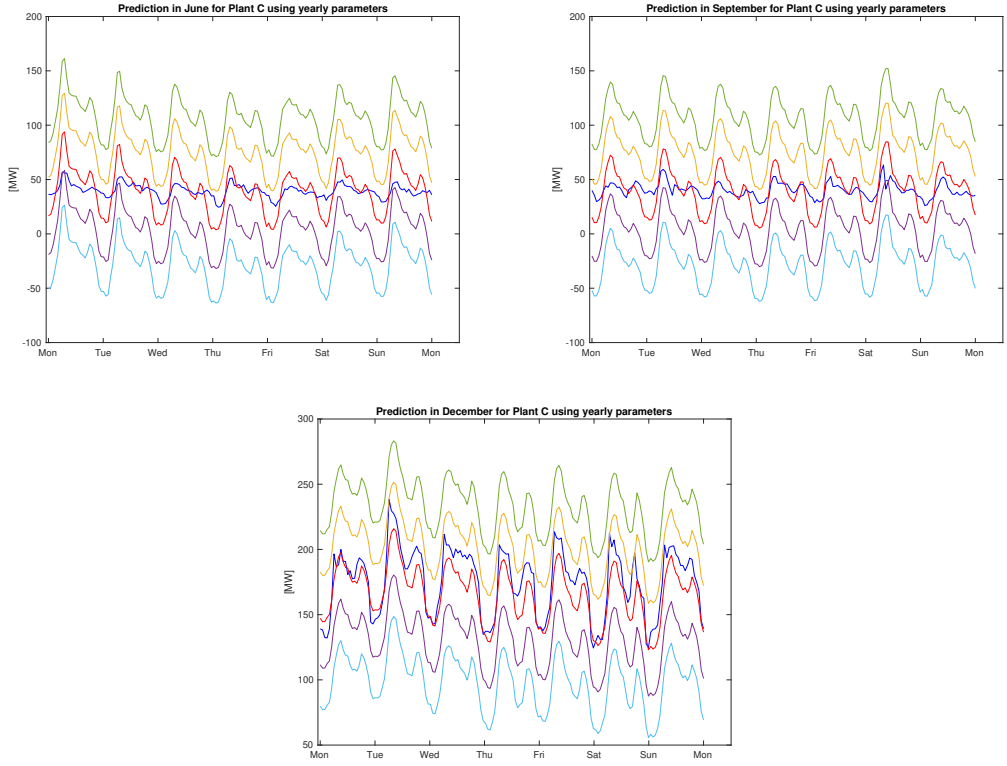


Figure 20: *Predicted heat load with confidence and prediction bands for plant C using yearly parameters. The periods shown are a week in June, September and December. The prediction in June and September is oscillating more than the true heat load and both bands' lower bound drops below 0 MW for some of the data points. In December the prediction is more accurate.*

4.2.2 Seasonal sets

Here follows the results of the seasonal predictions with and without confidence and prediction bands. We see in Figure 22 that the bands seem to get narrower for the July-September season. The prediction during summer in Figure 21 seems closer than for the yearly parameter model.

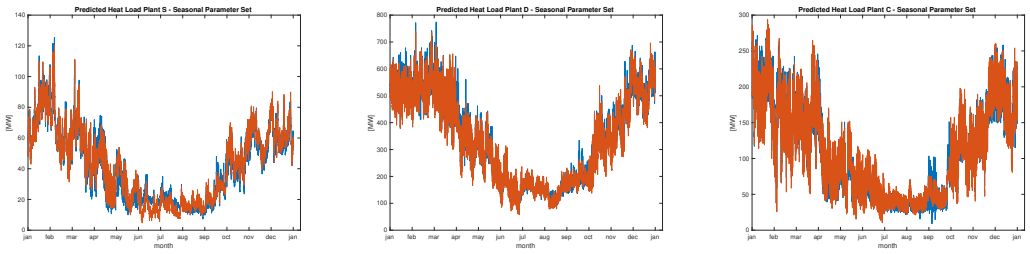


Figure 21: *Prediction and true DH for the seasonal parameter model for all plants.*

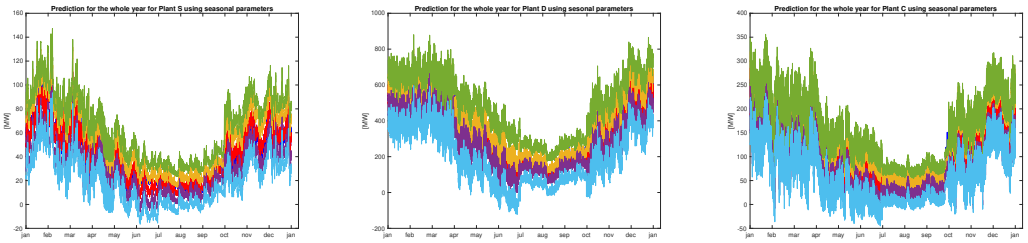


Figure 22: *Prediction and confidence bands added to the estimation and true DH for the seasonal parameter model for all plants.*

Table 8 below shows the size of the bands relative to the heat produced. The colder seasons have relatively narrower bands during the colder seasons, while April to June have the widest.

Table 8: *This table shows the size of the 95% confidence bands and 90% prediction bands in relation to the magnitude of the heat load for the seasonal partition models.*

(a) *Confidence level 95%*

Period \Plant	**S**	**D**	**C**
Jan-Mar	0.3096	0.1253	0.2056
Apr-Jun	0.3974	0.2320	0.5307
Jul-Sep	0.3610	0.1864	0.2619
Oct-Dec	0.2774	0.1956	0.3299

(b) *Prediction level 90%*

S	**D**	**C**
0.4054	0.2790	0.3798
0.7542	0.7127	0.8010
0.8220	0.6045	0.7751
0.4695	0.4109	0.3894

In Figure 23, 24 and 25 we can see the head load and corresponding bands for a week in June, September and December for all plants. Mostly the much narrower confidence bands seem to be able to contain the true hat load, however, not always for September. The prediction bands does seem to be wide enough for the studied weeks.

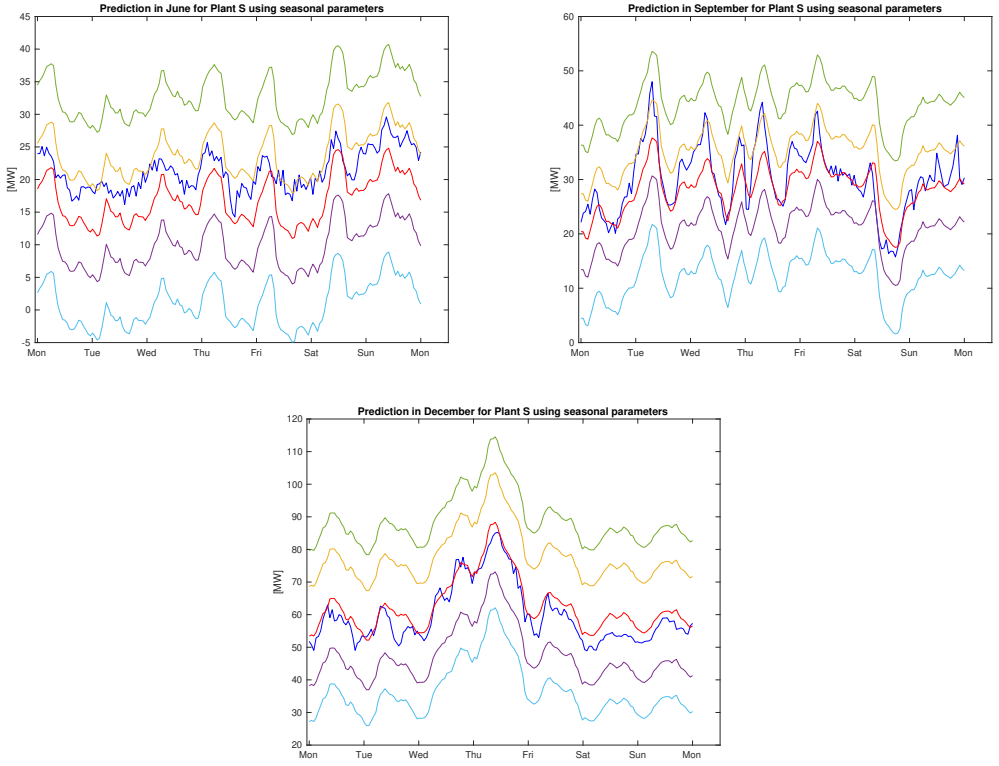


Figure 23: *Predicted heat load with confidence and prediction bands for plant S using seasonal parameters. The periods shown are a week in June, September and December.*

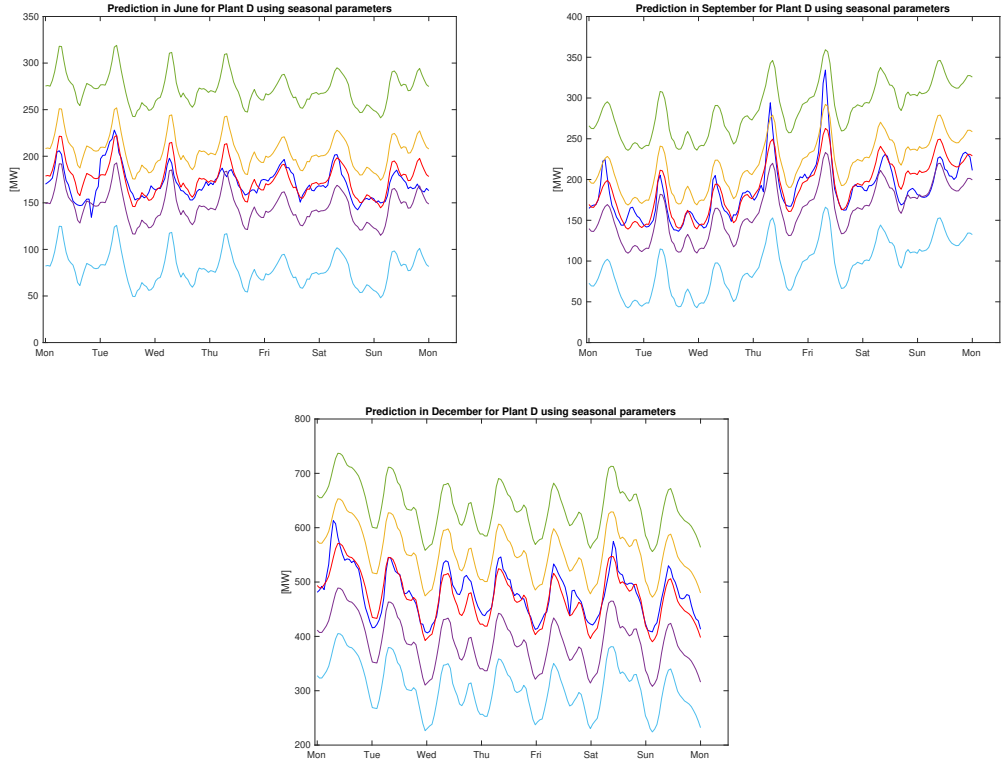


Figure 24: *Predicted heat load with confidence and prediction bands for plant D using seasonal parameters. The periods shown are a week in June, September and December.*

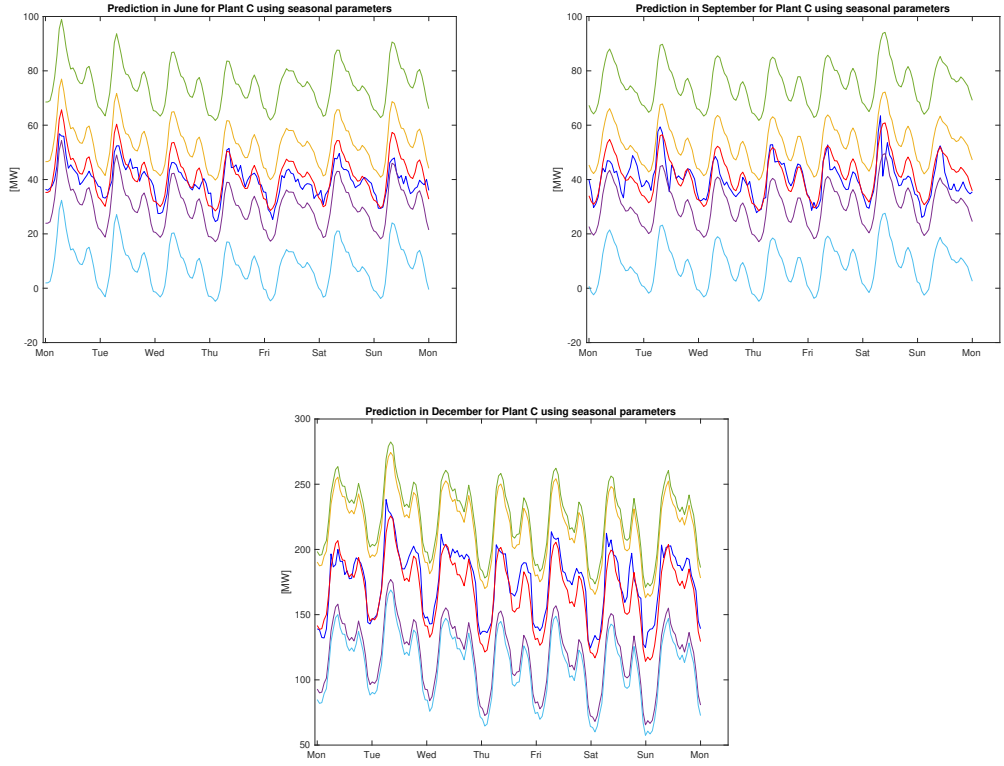


Figure 25: *Predicted heat load with confidence and prediction bands for plant C using seasonal parameters. The periods shown are a week in June, September and December.*

4.2.3 Monthly sets

Here follows the results of the monthly predictions with and without confidence and prediction bands. We see in Figure 27 that the bands seem to get narrower for the summer months. The prediction during summer in Figure 26 seems closer than for the yearly parameter model.

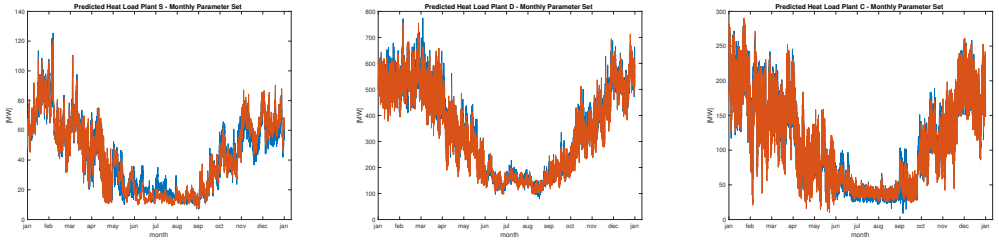


Figure 26: *Prediction and true heat load for all plants using monthly parameters.*

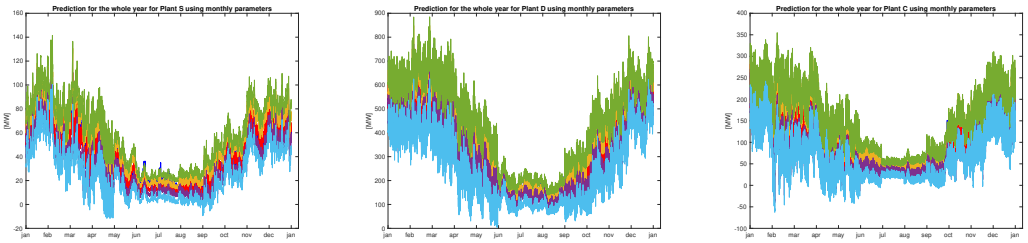


Figure 27: *Prediction and confidence bands added to the estimation and true DH for the monthly parameter set for all plants.*

Table 9 below shows the size of the bands relative to the heat produced for each month. January and December have relatively narrower bands, while months that are likely to appear during transition seasons, like April, May, August, September and October have wider bands. Remarkably, in contrast to the RMSPE scores, the bands during summer months are relatively narrow.

Table 9: *This table shows the size of the 95% confidence bands and 90% prediction band in relation to the magnitude of the heat load for the monthly partitions.*

(a) Confidence 95%

Period \Plant	S	D	C
Jan	0.1322	0.1191	0.0593
Feb	0.2223	0.1758	0.3868
Mar	0.3584	0.1408	0.4007
Apr	0.3679	0.1871	0.5053
May	0.3137	0.1885	0.5840
Jun	0.2106	0.1493	0.2272
Jul	0.1669	0.0990	0.1377
Aug	0.3768	0.1161	0.1768
Sep	0.3637	0.1992	0.2554
Oct	0.2397	0.2509	0.3633
Nov	0.1951	0.1736	0.2168
Dec	0.1491	0.0965	0.1727

(b) prediction 90%

S	D	C
0.2367	0.1947	0.2476
0.3171	0.2394	0.5325
0.4579	0.2744	0.4689
0.5892	0.3886	0.6246
0.5661	0.4749	0.7410
0.4559	0.3193	0.7019
0.4544	0.3356	0.3830
0.7937	0.3099	0.6148
0.6632	0.4761	0.5739
0.3461	0.4031	0.3661
0.3429	0.2649	0.3190
0.3116	0.1686	0.2573

In Figure 28, 29 and 30 we can see the head load and corresponding bands for a week in June, September and December for all plants. Mostly the much narrower confidence bands seem to be able to contain the true hat load, however, not always for June and September. Especially plant S have a prediction that is clearly lower than the true heat load in June. The prediction bands does seem to be wide enough for the studied weeks.

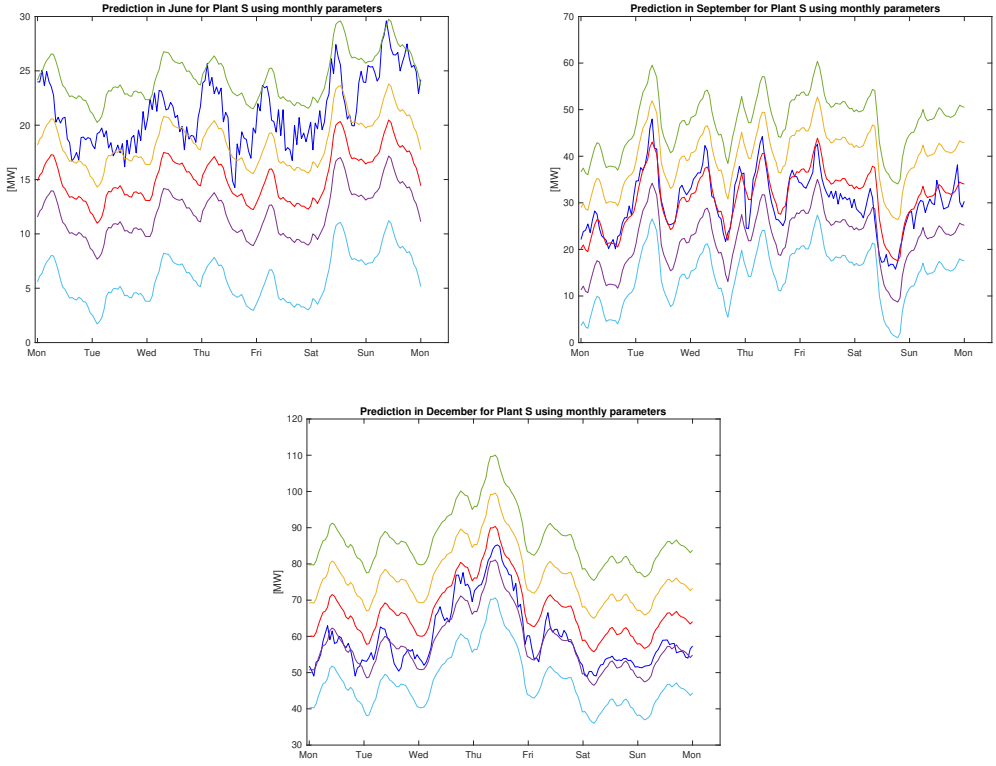


Figure 28: *Predicted heat load with confidence and prediction bands for plant S using monthly parameters. The periods shown are a week in June, September and December. The prediction in June is slightly low and almost go outside the prediction band as well. In December it is the opposite, while September is more centred.*

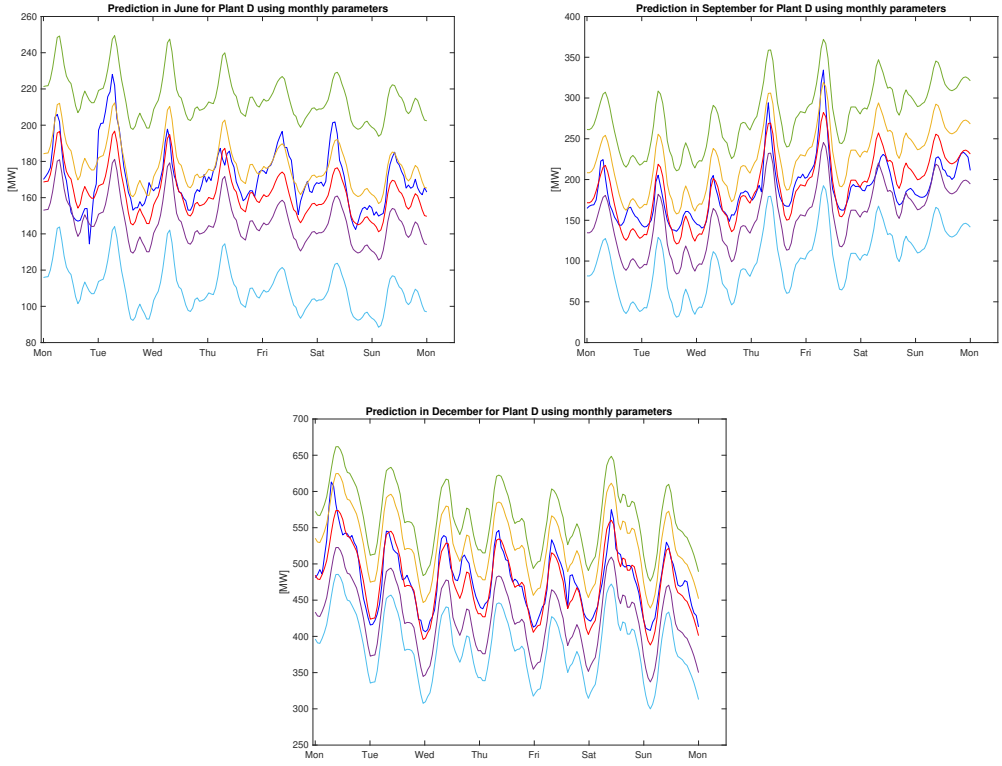


Figure 29: *Predicted heat load with confidence and prediction bands for plant D using monthly parameters. The periods shown are a week in June, September and December. The prediction in June misses some peaks, which go outside the confidence band but is contained within the prediction band. September and December have more narrow bands which still contain the model prediction within most of the confidence band.*

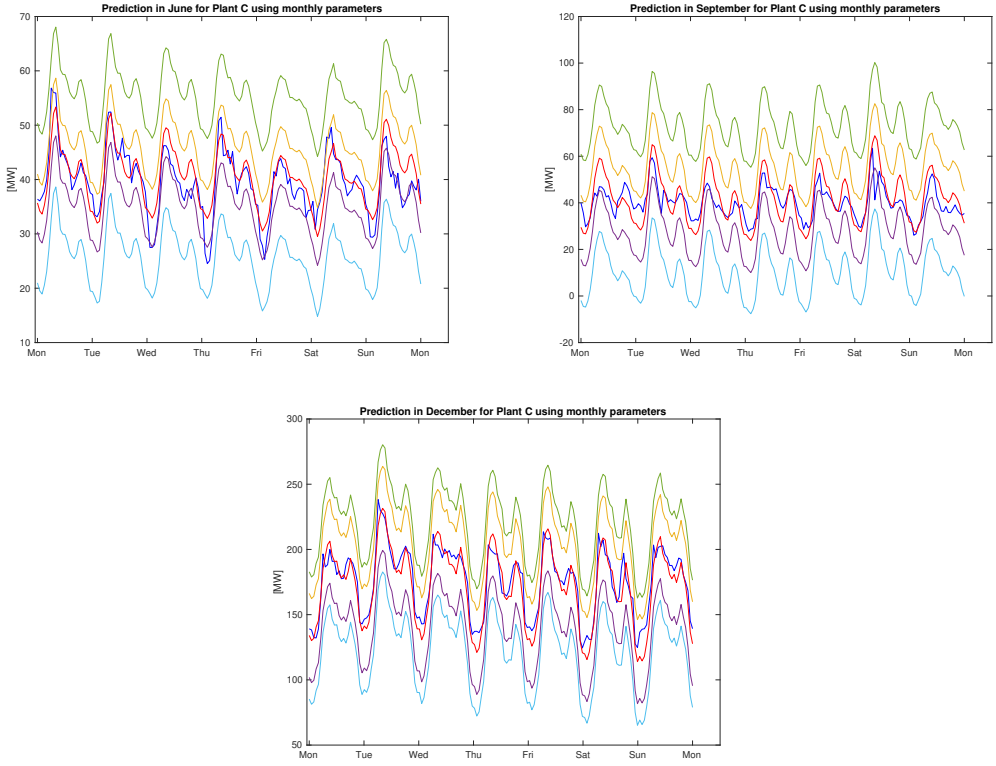


Figure 30: *Predicted heat load with confidence and prediction bands for plant C using monthly parameters. The periods shown are a week in June, September and December. Overall the prediction looks good and the true heat load is even contained within the confidence band for the most part. The confidence and prediction bands seem to be similar in size for December, and greater in June.*

4.2.4 Weekly sets

Here follows the results of the weekly predictions with and without confidence and prediction bands. The predictions in Figure 31 seem to follow the general shape of the heat load, but does have some odd spikes for some weeks. We see in Figure 32 that the bands seem to get narrower for the summer months for plant D and C. For plant S, however, the bands are vary wide at some points. All plants have some intervals that get much wider bands than the others'.

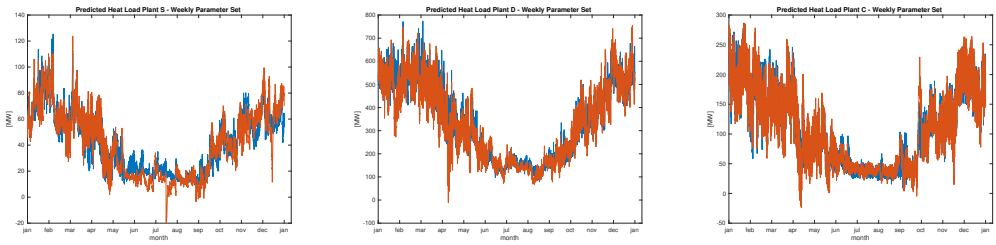


Figure 31: *Prediction and true heat load for the model with the weekly parameter set for all plants*

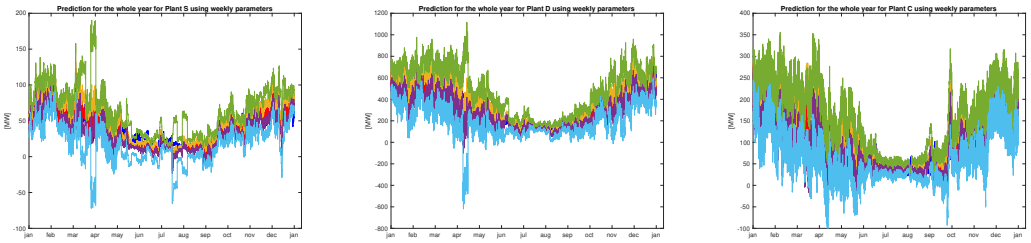


Figure 32: *Prediction and confidence bands added to the estimation and true DH for the weekly parameter set for all plants. There are varying sizes of prediction bands from week to week.*

No table was made for every single week, but closer looks at a week in June, September and August was made. As can be seen in Figure 33, the weekly prediction is way off in both June and September for plant S, not even the prediction bands contain the true heat load. While plant D and C, in Figure 34 and 35 respectively, seem to have a better prediction than plant S, the confidence band does not contain the true heat load all the time, but the prediction bands does. We can also see when the parameters change and the bands changes a lot in size for June and December for plant D, as seen in Figure 34.

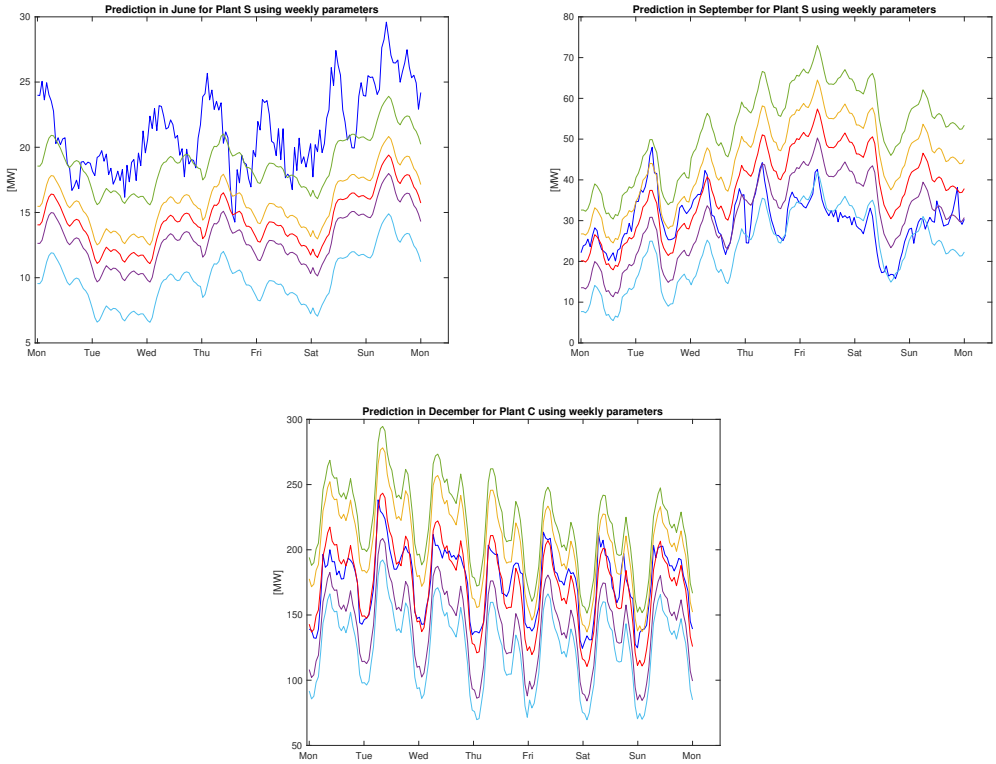


Figure 33: *Predicted heat load with confidence and prediction bands for plant S using weekly parameters. The periods shown are a week in June, September and December.*

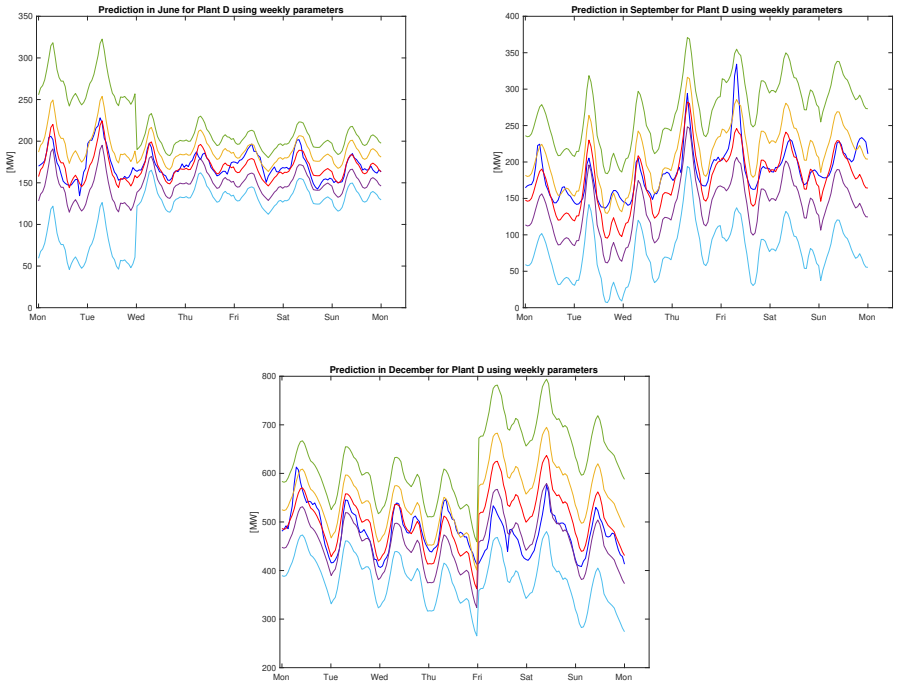


Figure 34: *Predicted heat load with confidence and prediction bands for plant D using weekly parameters. The periods shown are a week in June, September and December.*

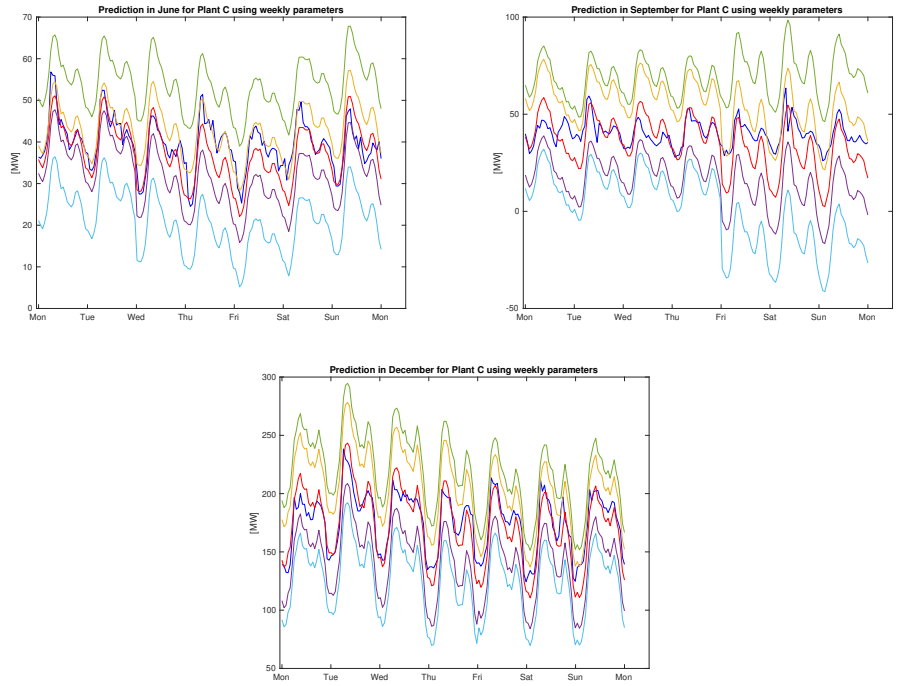


Figure 35: *Predicted heat load with confidence and prediction bands for plant C using weekly parameters. The periods shown are a week in June, September and December.*

5 Discussion

For this sections, the results will be discussed and answers to the objectives sought after. There will also follow a discussion about sources of errors and further studies that could be conducted based on the results found here.

5.1 Analysing the results

Having some partitions, but not too many seems to be better for prediction. When it comes to fitting the model, the more partitions used, the better the model fits to the measured data as seen in Table 2. This adds up, as the model can fit better to each part. More importantly however, is what happens when the parameters are used for prediction. In Table 3 we can see that that the yearly and weekly parameter sets performed worst for all three plants. These results indicate that a dividing the year into shorter seasons, at least all the way down to a month is likely to perform better that one that uses the same parameters throughout the whole year. On the other hand, too small intervals may lead to the fitting being too specified and the variability for that period may be larger than acceptable.

The yearly set could still be acceptable if one does not have the computational means of using multiple sets. As long as one are aware that it probably will perform poorly during warmer seasons. If it is possible to turn off the temperature affected parameters completely for the warmer periods, it could possibly be perform acceptable during summer as well.

A purely weekly parameter set is probably not the best solution for predicting, however. Just looking at the graphs in Figure 32, the prediction bands are very large for some weeks. The prediction itself seem to also vary quite a bit for when it predicts well and when it does not. So even if it may have some accurate intervals, the unpredictability is not desirable for a reliable prediction.

The difference in weather may also affect the accuracy of the model depending on which parameter set is used. When looking at the prediction in a week in June for plant S, we see in Figure 18,23, 28 and 33, the monthly set predicted less than the actual heat load needed, while the weekly predicted even less and the seasonal and yearly predicted closer to the true heat load, albeit with stronger oscillations. In Figure 11 we can see that the heat load in 2018 was a bit lower during summer than it was in 2019. So even if the temperature would be lower for the plotted week, if the model has set a base line that is too far below the heat load of a predicted year, the temperature might not be able to weigh up against it depending on the parameters set for temperature values. The yearly and seasonal parameters are not as affected by this offset for a period in the summer as they fit the model during a longer period of time. The trade off here is then an offset versus oscillations.

The time of the year with largest RMSPE seem to be around summer and the transition periods to and from summer. This is a bit surprising, as prediction during

summer usually goes well from the experience of the Company. However, the plants that had a low RMSPE for summer were plant S and D. Looking at June in Figure 28 and 29, the heat load is consistently higher than the prediction, especially for Plant S. This is possibly because the heat load was lower the period before, which the model then had a hard time adjusting for. Although, the confidence and prediction bands during summer are not particularly wide, the prediction band contains the true heat load, even if the confidence band does not. It might be that summer is usually easy to fit the model to, which may lead to a worse prediction if there is an offset for next year, either because the prediction oscillates with the temperature more, or because it fit too closely to last year. In general when choosing a parameter set in regards to summer, all three plants got the best RMSPE score when using monthly parameters for June to August, with the exception of plant S in July, where seasonal parameters had a slight edge.

The transition periods are more likely to vary more, however. Only plant C had the highest RMSPE during a transition month, September. It may have been affected by the strong oscillations occurring in the beginning of that month, as can be seen in Figure 10. Regardless, looking at Table 9, the month with the largest confidence and prediction bands all lie in a possible season transitional period. So while the overall relative RMSPE may be lower, there is likely the highest variability in heat load during those periods. The prediction bands turn large, but in return they do seem to cover the unforeseen peaks as seen in the September graphs in Figure 29 and 23. As for choosing a parameter set for these periods, the monthly variant is not always better than the seasonal. It is hard to tell when the weather is unpredictable as well and perhaps going for the model with the smaller prediction band gives the most reliable forecast.

Except for the weekly model, all models were able to produce a relatively accurate model during the colder seasons. Both the RMSPE and relative size of the confidence and prediction bands are quite low during November to February. Not only is the heat load largest during this period, but looking at December in any figure using yearly, seasonal or monthly parameters depicting it, the prediction does seem to follow the true heat load quite closely. From the RMSPE values all three of those partitions can work well during winter times.

All in all, the confidence and prediction bands seem to reliably contain the future heat load. While the prediction bands do grow quite large, even peaks in transitions seasons as mentioned above are contained within. Since the point of the prediction bands are to contain the whole regression line for future values, they have to be large for periods that have large variations in heat load demand. The bootstrapped prediction bands also seem to be able to pick out when the input values are differing between the years. Take the summer months as an example. In Table 9 is about two to three times larger than the confidence bands for the summer months for all plants, even larger for plant C. While growing in size, they do manage to cover the odd peaks in summer that shows up. However, the confidence bands seem to be reliable over all as well. There are some peaks in the transitions months that do get missed

and there are intervals during summer that escape the confidence confinements, but overall, the true heat load stays inside most of the time. In conclusion, the confidence bands work well for prediction, but if the prediction bands turn out much larger, there may be a large difference in input data compared to last year.

Weather aspects other than temperature may be helpful for a more robust prediction. Looking both at the RMSPE in Table 5 and the the values for Plant D in Table 9 in comparison for the corresponding tables, Table 4, 6 and values in Table 9 for plant S and C, plant D in general have proportionally narrower bands and a lower RMSPE. As plant D is the only one of the three to additionally use wind speed and solar influx as input to the model, they could possibly affect the model in a positive way. On the other hand, plant D is the plant with the largest production, so the sheer magnitude of heat load produced may make it more robust against smaller differences between the years in weather input. This is probably the more likely scenario, as an earlier study and in the experience of the company, this extra input data usually does not affect the prediction significantly [5]. Then again, the geographic location of the plants may have some effect. Since plant D is located in Denmark, which is relatively flat, the wind is likely to be stronger than in more rugged area.

5.2 Sources of Errors

In reality, the weather measurements are also estimated when doing prediction. We do not know exactly what the temperature will be tomorrow and have to rely on weather forecasts. In this study, we used already known data for prediction analysis and with the more realistic scenario, the prediction would have an additional factor of uncertainty added. This could potentially make the summer periods relatively easier to predict, in the case that weather during the summer is more predictable. This may also have affected the prediction during transition periods between seasons like winter to spring, usually occurring in April, less reliable as we know from experience how the April weather can be. In short, the RMSPE and size of the bands may shift around more depending on uncertain weather data.

Some measurements both in weather data and heat load may have been wrongly measured. It happens to most technical apparatuses that they sometimes get a wrong reading or are turned off. Usually the company tries to manually adjust for any missing or out of place measurements, however there may be some singularities and if many days are missing, the data can be a bit misleading.

For Plant D, we removed a day in July that were missing data. This may pose a bit of a misleading prediction as the weekdays for the rest of the years then got shifted data-wise, but not model-wise. This means that the model might predict a Saturday as if it were a Friday instead and since the model contain parts regarding social patterns, that differs depending if it is a weekday or not, the results may be somewhat shifted. I noted this a bit late and did not have time to adjust it.

5.3 Further Studies

Choosing block length for the MBB could possibly have been performed better. As mention in Section 2, choosing the length to be only 24 there may be loss in information regarding the day of week and all that entails for the heat load. There is also much more to read up on regarding choosing a suitable block length, which was not done thoroughly for this project, as I simply picked one that would fit for all interval lengths dealt with in this study. An idea could be to choose different block lengths depending on the number of days in the interval, though it requires an extra bit of programming and reasoning and may also become harder to compare the results for each partition.

In this study, all partitions were of similar length throughout the year, but does not necessary need to be the case. Perhaps it would also be wiser to let the seasonal parameters represent the actual seasons, rather than strict quarters of the year. Those could also be based on meteorological seasons, which would require weather forecasts. Or in general base the chosen parameter set on weather forecast. Of course this would put one at mercy of having a good weather forecast, but it could be worth looking into.

For this study, linear regression with standard least squared error was used, however, that may not be the optimal regression method for heat load. Since undershooting in winter or over shooting in summer generally is seen as worse than the opposite, it might have been better to use an asymmetric least squared (ALS). ALS gives weights to the errors, so missing peaks in winter would be regarded as more wrong than overshooting. It would be interesting to see if a monthly or seasonal approach still would yield a better prediction when the errors are weighted differently.

This study focused more on the relative sizes of the bands and used RMSPE as a tool for analysis. So even if the summer is shown to have the largest RSMPE for most plants, the losses during summer might not matter particularly much compared to the errors in winter, even if they are smaller relatively speaking. The results of this study are not to be disregarded because of that, they are still of interest, however, looking at more absolute values would also be worth examining.

6 Conclusions

To summarise what was gained from this study, using a few partitions throughout the year will yield a more accurate prediction. Using the same parameters all year will especially be less reliable during summer, while using too many partitions, like weekly, will result in the forecast having a greater variability. Whenever the input data differs much between the fitted year and the predicted one, the variants with smaller partitions are more susceptible to offsets, while those with larger might give more oscillations.

For this data set, summer was generally the season with highest RMSPE for the predictions, however, using monthly parameters seems to give the better forecast for all plants. The transitions times between spring, summer and autumn seem to be the time with most variability in the data, which the bootstrapped bands pick up on. Both seasonal and monthly performed nicely during those times. All partitions but the weekly can quite reliably predict the heat load during winter time.

In general, the bootstrapped bands keep the true future heat load within their bounds, and even in periods where the confidence band fails, such as for offsets and large peaks, the prediction band keeps up, although being relatively large. Either some weather inputs additional to temperature, or having a larger heat production, may help the stability of the forecast. All in all, using partitions somewhere between one to three months seem to give a more reliable and accurate prediction.

References

[1] S. Frederiksen and S. Werner. *Fjärrvärme: teori, teknik och funktion*. Studentlitteratur, 1993.

[2] S. Werner and H. Averfalk. *Essential improvements in future district heating systems*. 2017. URL: https://www.diva-portal.org/smash/get/diva2:1153175/FULLTEXT01.pdf (visited on 12/02/2021).

[3] C. Johansson. *ON INTELLIGENT DISTRICT HEATING*. 2014. URL: http://www.diva-portal.org/smash/get/diva2:834222/FULLTEXT01.pdf (visited on 08/03/2021).

[4] Euroheat. *District Energy in Sweden*. Nov. 2019. URL: https://www.euroheat.org/knowledge-hub/district-energy-sweden/ (visited on 10/02/2021).

[5] J. Henriksson and S. Rudén. *Investigation of the Improvement Potential of Heat Load Forecasts in BoFiT*. 2018. URL: https://www.diva-portal.org/smash/get/diva2:1295909/FULLTEXT01.pdf (visited on 02/03/2021).

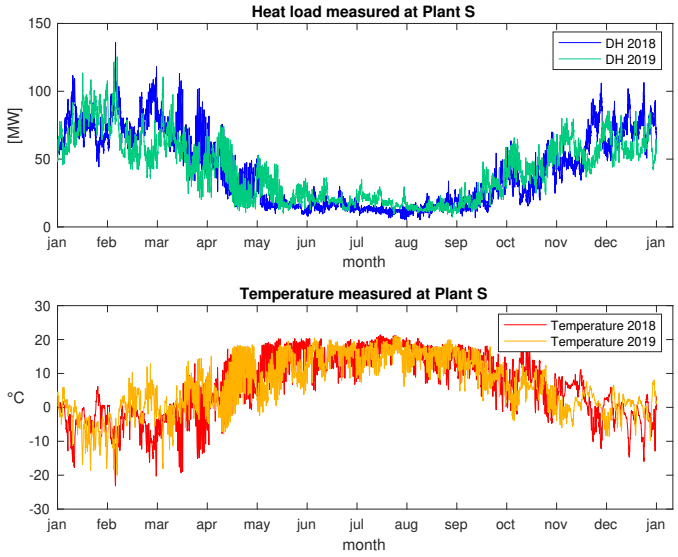
[6] J.S.Urban Hjort. *Computer Intensive Statistical Methods*. Chapman & Hall, 1994. ISBN: 0 412 49160 5.

[7] F. Bergström. *Bootstrap Methods in Time Series Analysis*. June 2018. URL: https://kurser.math.su.se/pluginfile.php/20130/mod_folder/content/0/Kandidat/2018/2018_4_report.pdf?forcedownload=1 (visited on 30/03/2021).

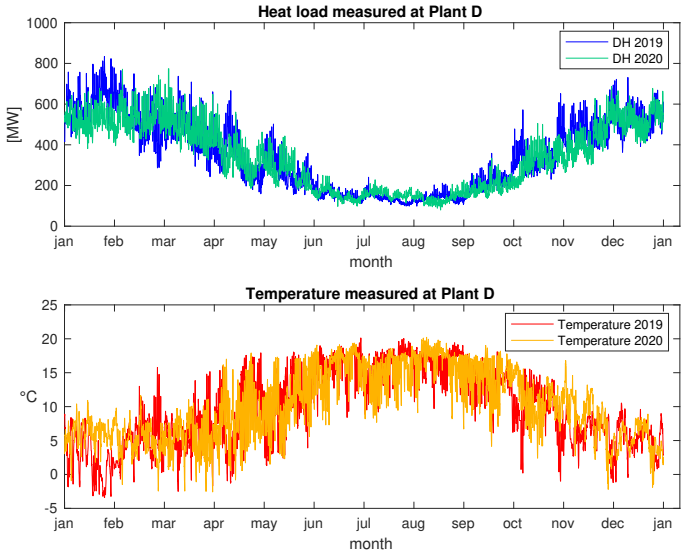
[8] Robert A. Stine. 'Bootstrap Prediction Intervals for Regression'. In: *Journal of the American Statistical Association* 80.392 (1985), pp. 1026–1031. ISSN: 01621459. URL: http://www.jstor.org/stable/2288570 (visited on 10/05/2021).

A Heat load and Temperatures

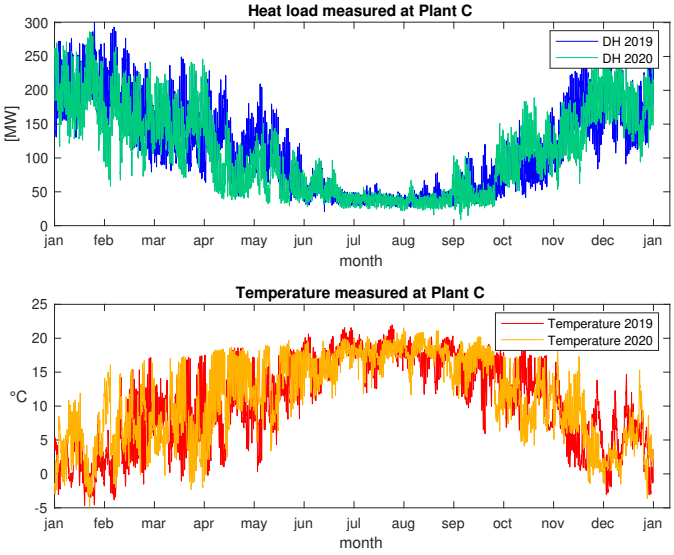
Here the heat load and temperature for both the fitting year and the prediction year are plotted together in higher resolution.



This depicts the district heat load and temperature for both the year used for fitting the model and for predicting for Plant S.



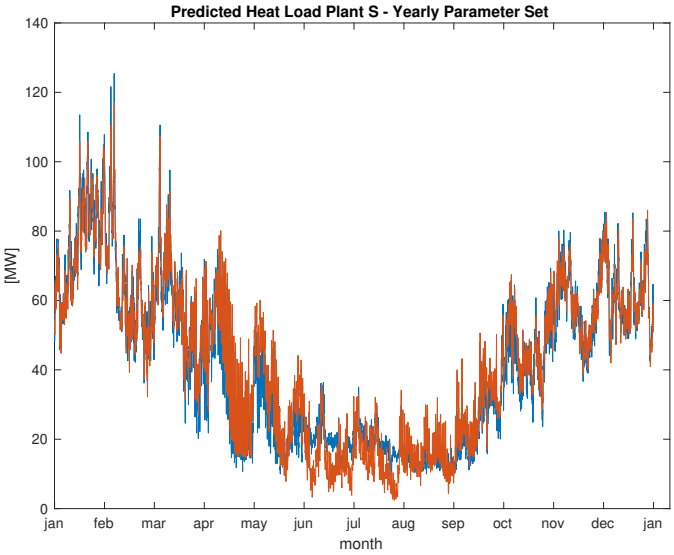
This depicts the district heat load and temperature for both the year used for fitting the model and for predicting for Plant D.



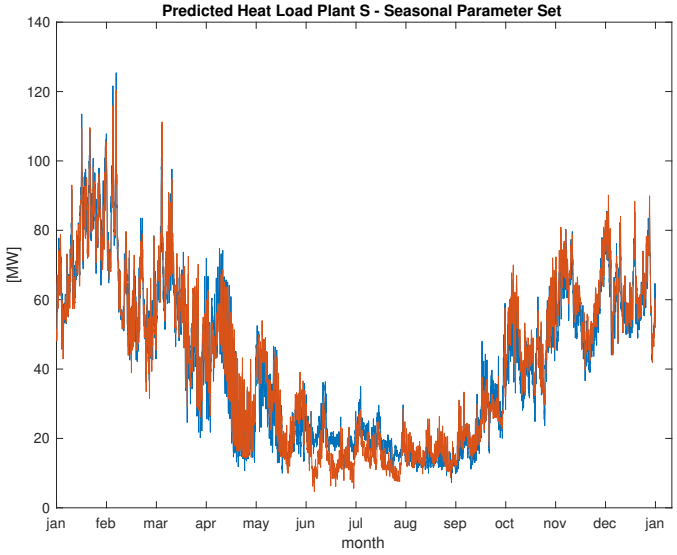
This depicts the district heat load and temperature for both the year used for fitting the model and for predicting for Plant C.

B Model Predictions

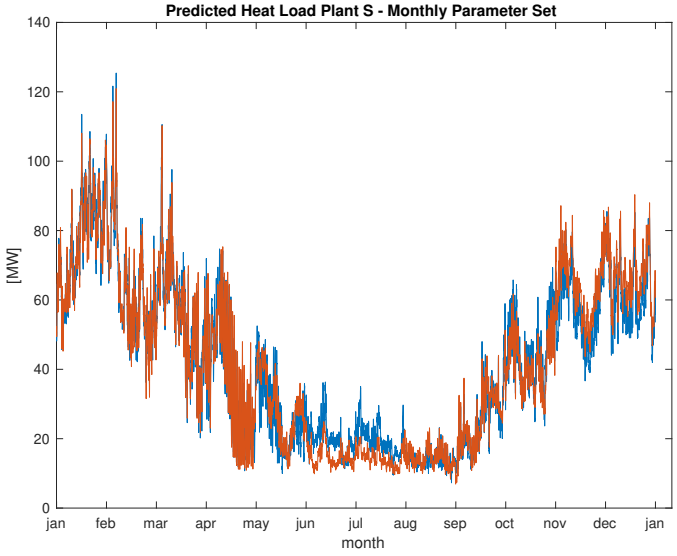
This part shows the model predictions in higher resolution. First all predictions for plant S will be shown, then plant D and finally plant C



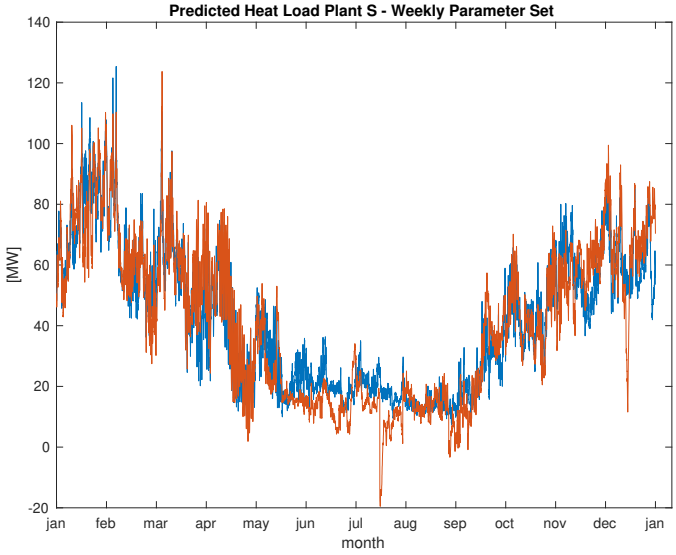
Prediction and true heat load for the model with the yearly parameter set for plant S



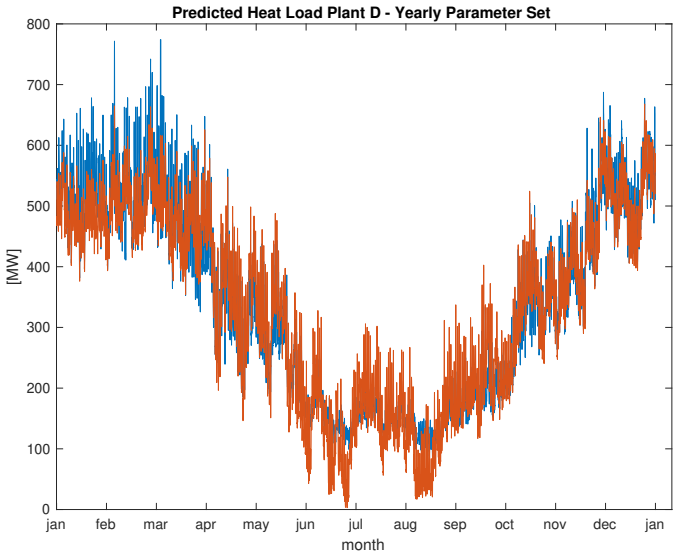
Prediction and true heat load for the model with the seasonal parameter set for plant S



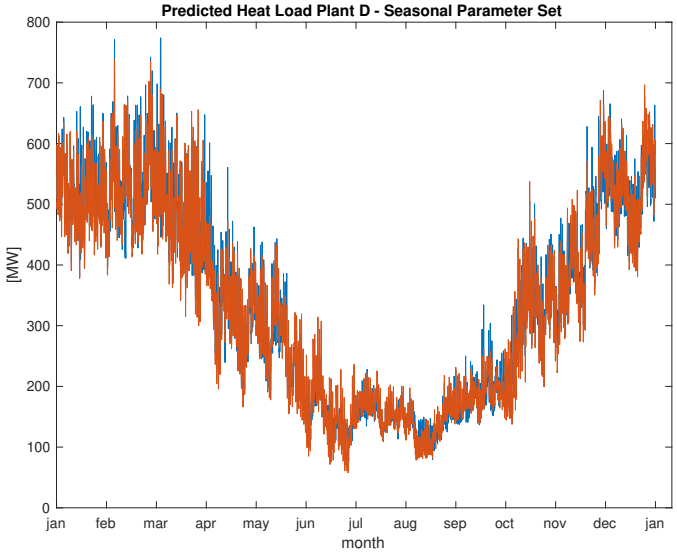
Prediction and true heat load for the model with the monthly parameter set for plant S



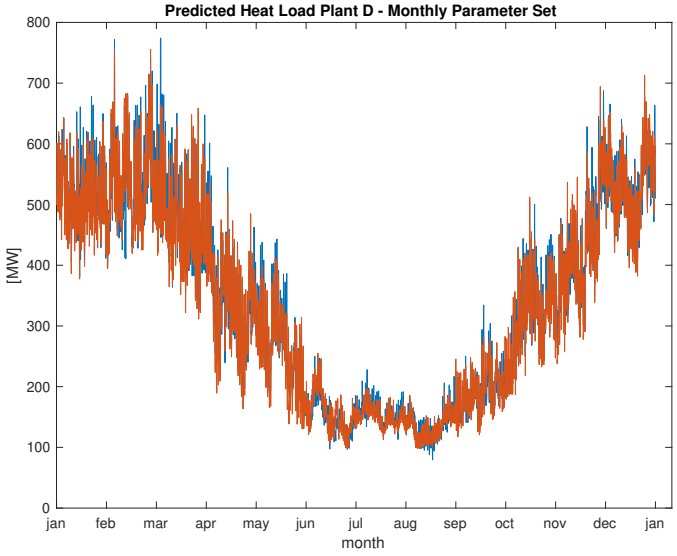
Prediction and true heat load for the model with the weekly parameter set for plant S



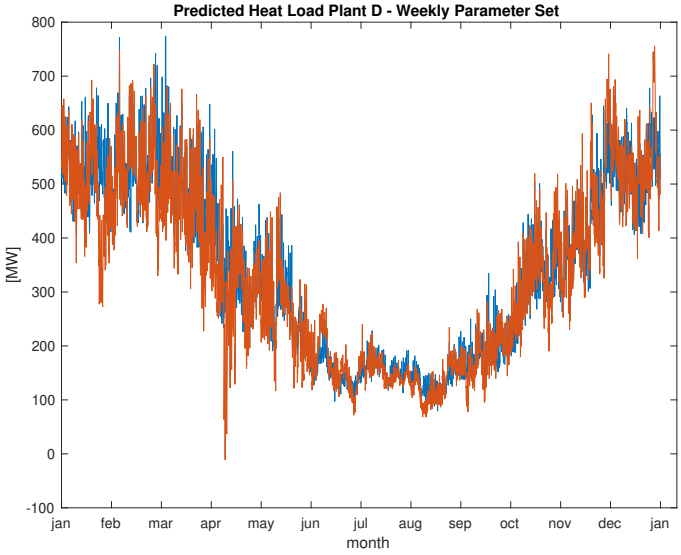
Prediction and true heat load for the model with the yearly parameter set for plant D



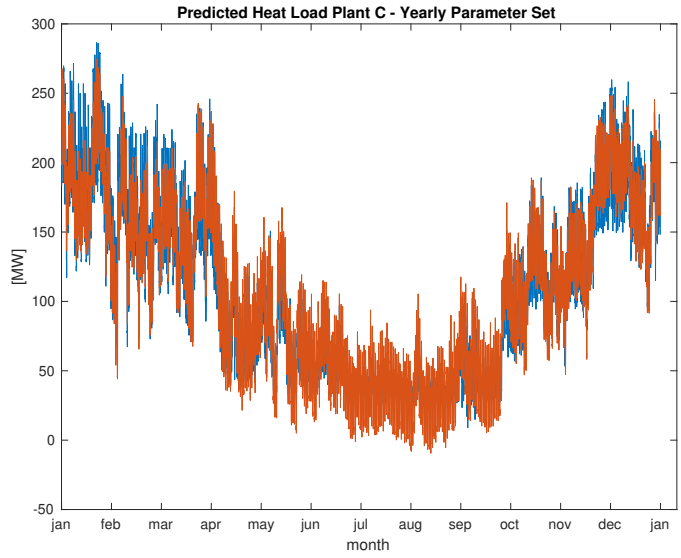
Prediction and true heat load for the model with the seasonal parameter set for plant D



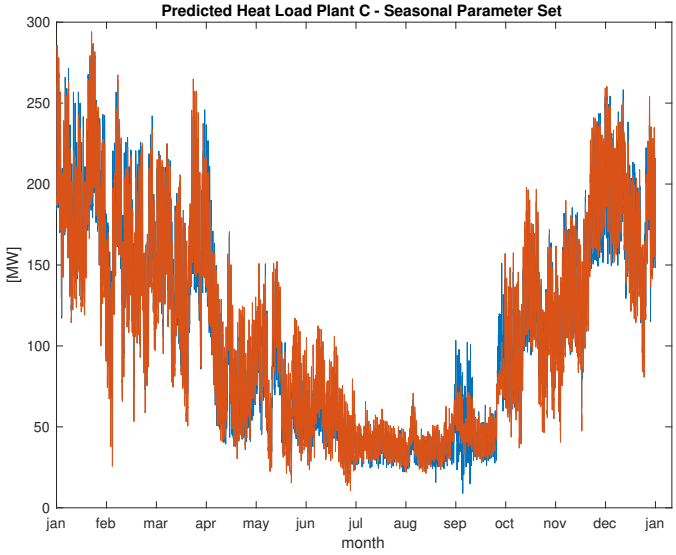
Prediction and true heat load for the model with the monthly parameter set for plant D



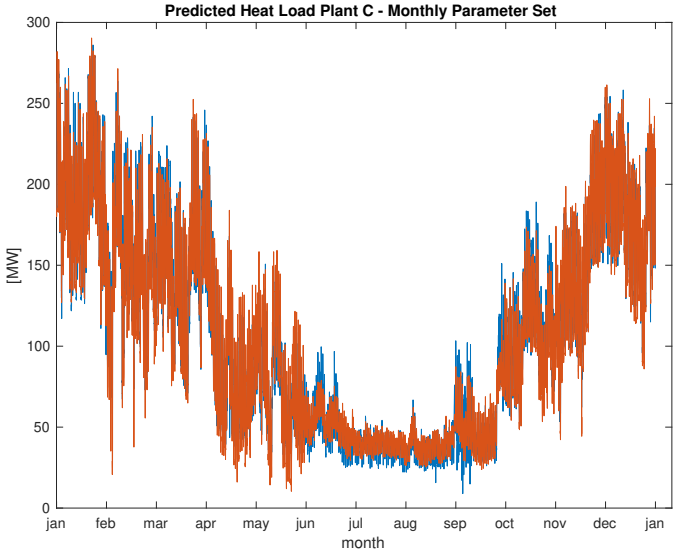
Prediction and true heat load for the model with the weekly parameter set for plant D



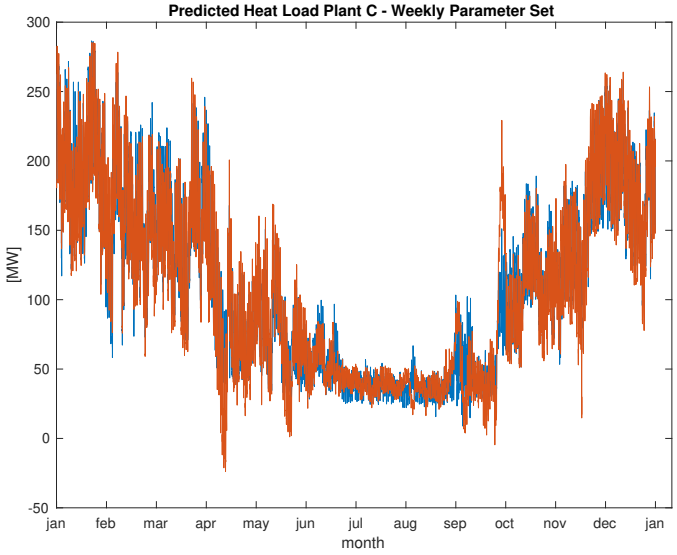
Prediction and true heat load for the model with the yearly parameter set for plant C



Prediction and true heat load for the model with the seasonal parameter set for plant C



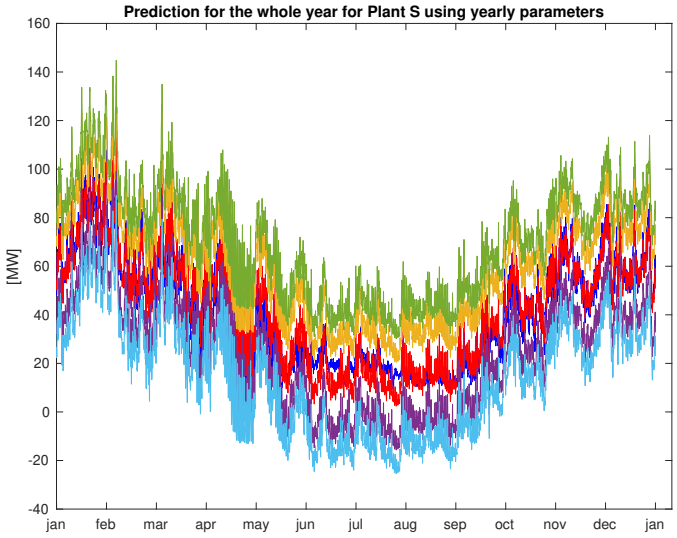
Prediction and true heat load for the model with the monthly parameter set for plant C



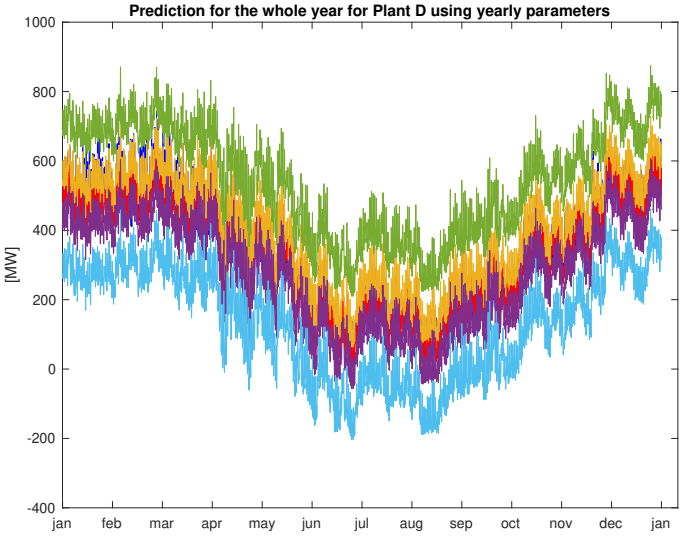
Prediction and true heat load for the model with the weekly parameter set for plant C

C Confidence and Prediction Bands

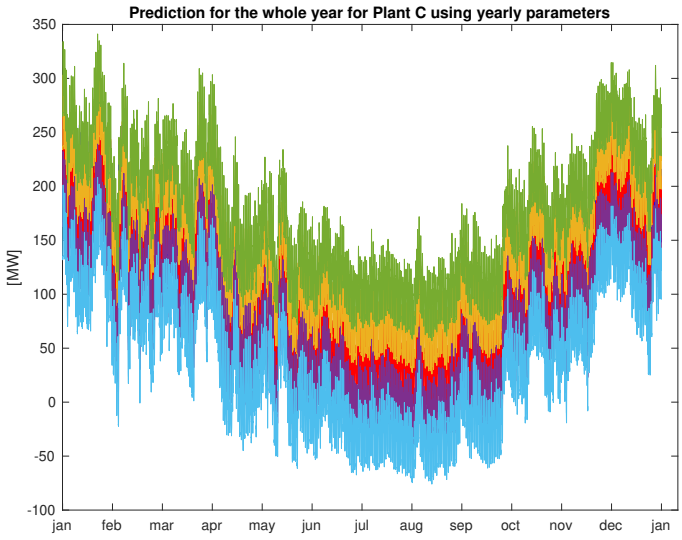
The confidence and prediction bands of the whole year are shown for all plants in greater resolution. First plant S, then plant D and lastly plant C.



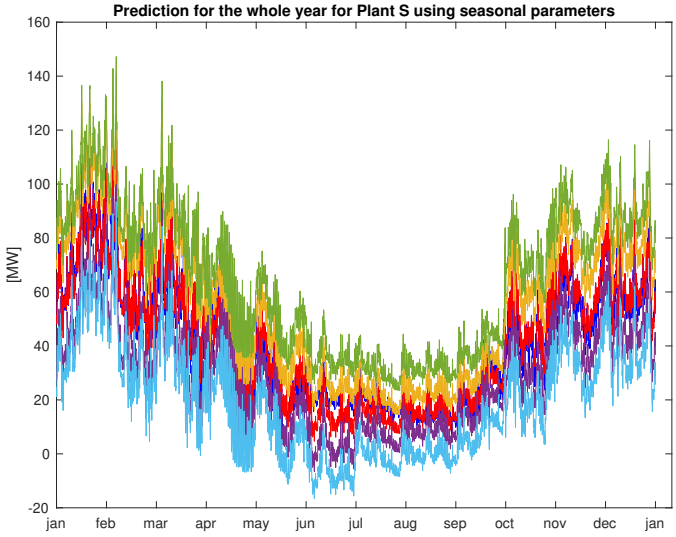
Prediction and confidence bands added to the estimation and true DH for the yearly parameter set for plant S.



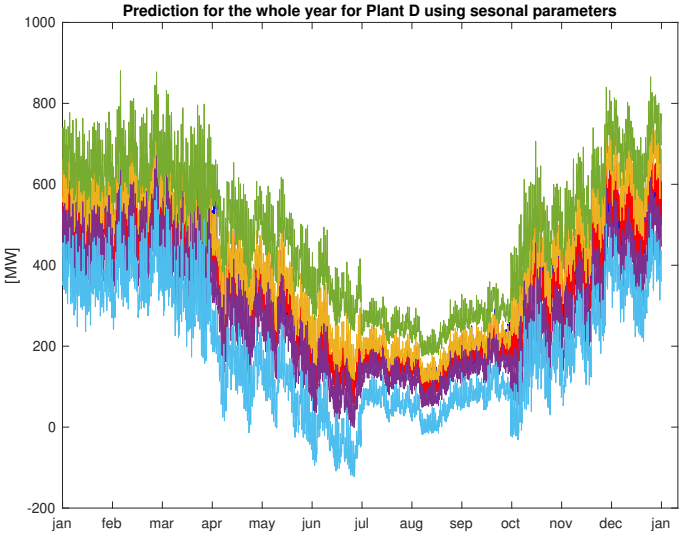
Prediction and confidence bands added to the estimation and true DH for the yearly parameter set for plant D.



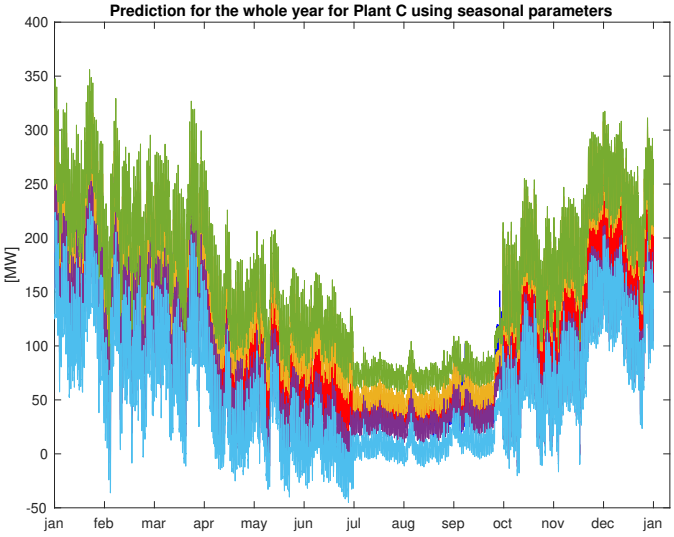
Prediction and confidence bands added to the estimation and true DH for the yearly parameter set for plant C.



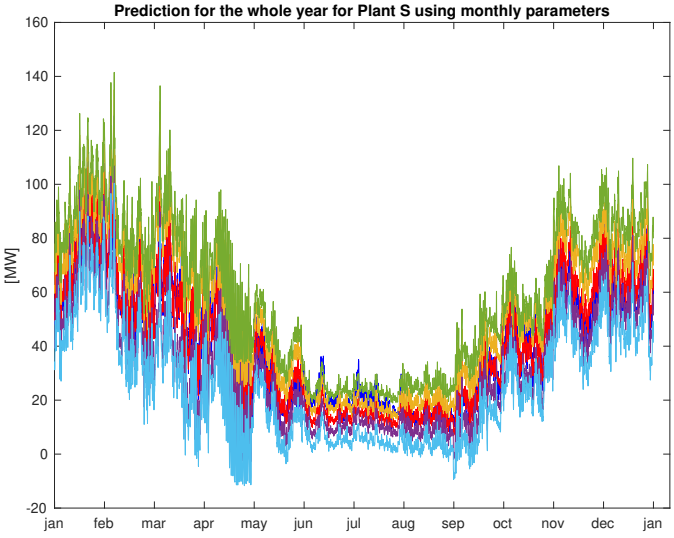
Prediction and confidence bands added to the estimation and true DH for the seasonal parameter set for plant S.



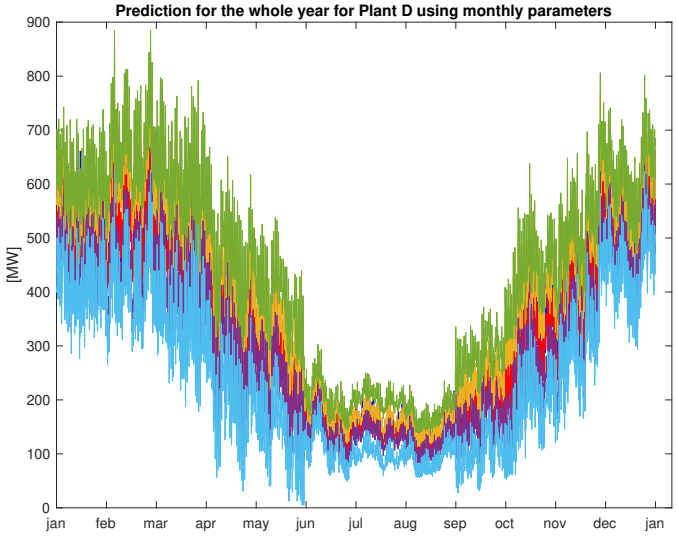
Prediction and confidence bands added to the estimation and true DH for the seasonal parameter set for plant D.



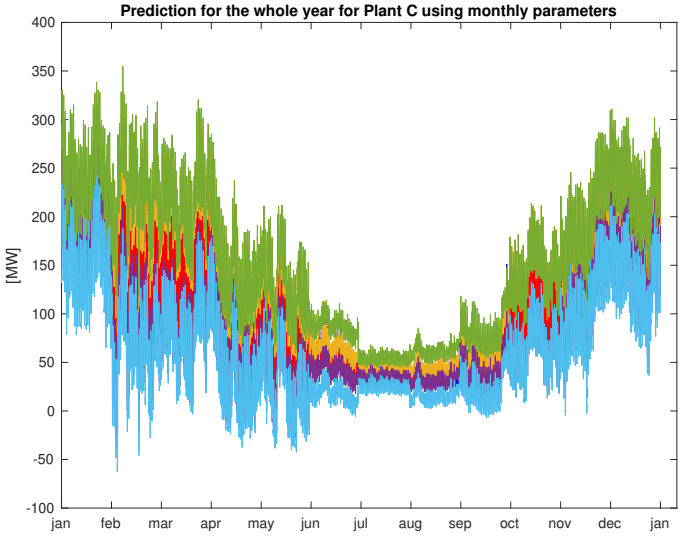
Prediction and confidence bands added to the estimation and true DH for the seasonal parameter set for plant C.



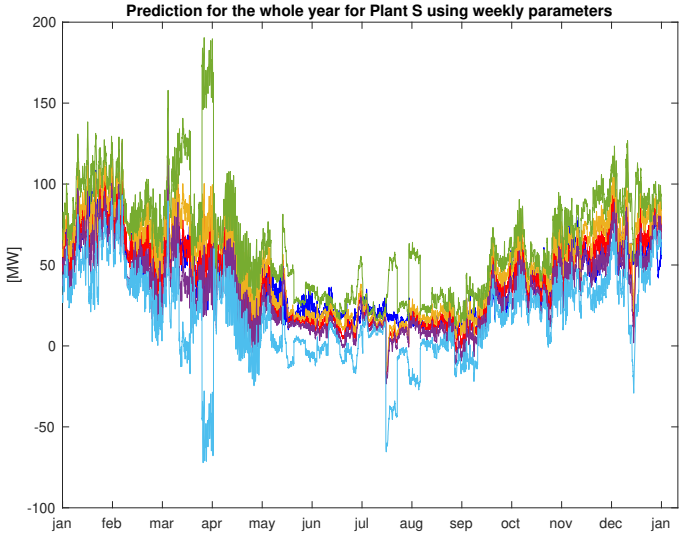
Prediction and confidence bands added to the estimation and true DH for the monthly parameter set for plant S.



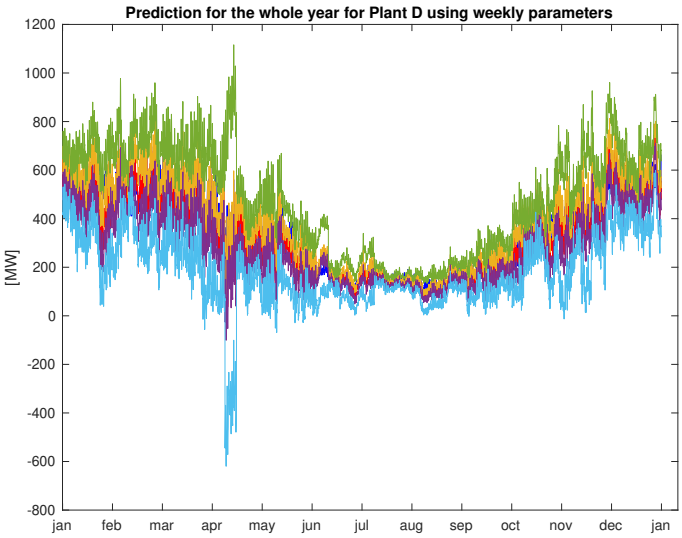
Prediction and confidence bands added to the estimation and true DH for the monthly parameter set for plant D.



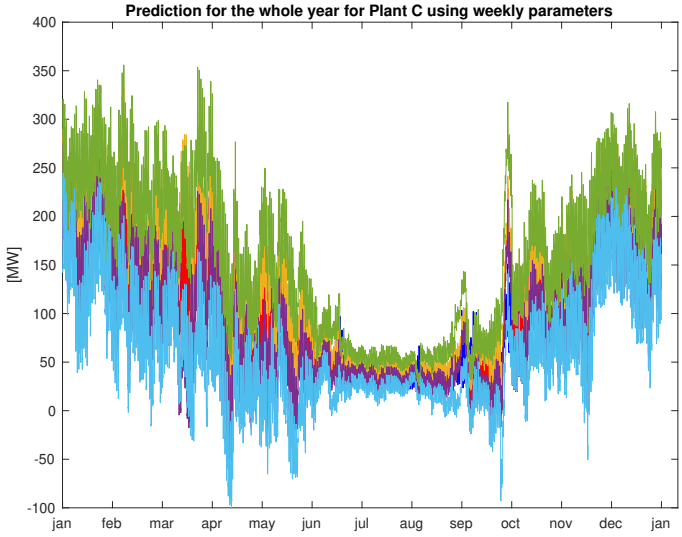
Prediction and confidence bands added to the estimation and true DH for the monthly parameter set for plant C.



Prediction and confidence bands added to the estimation and true DH for the weekly parameter set for plant S.



Prediction and confidence bands added to the estimation and true DH for the weekly parameter set for plant D.



Prediction and confidence bands added to the estimation and true DH for the weekly parameter set for plant C.