BACHELOR

# BEP Predictions Dutch Storm Barrier

Michels, Joshwa R.

*Award date:*
2020

Technische Universiteit Eindhoven

# BEP Predictions Dutch Storm Barrier

Joshwa Michels
November 15, 2020

# Contents

# Chapter 1

# Context

The Dutch Ministry of Infrastructure and Water Management (Rijkswaterstaat in Dutch) constructed storm surge barriers to provide protection against floods. The Eastern Scheldt storm surge barrier (Oosterscheldekering in Dutch) is one such barriers. It was built by taking into consideration the effect it has on nature. In particular, the region is a home for many aquatic animals especially for mussels and oysters.

The Eastern Scheldt Storm Surge Barrier is supposed to be closed whenever there is an emergency that could bring damage to nature and the community. By law, the barrier should be closed only when the water is exactly 3.00 metres above sea level at the measurement location RPBU (see Figure 1.2). Closing the sluices of the storm surge barrier takes about 30 minutes(see Figure 1.1).

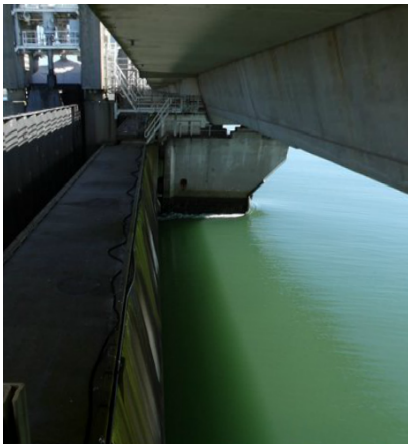

Figure 1.1: Side view of one of the sluices of the Eastern Scheldt Barrier, taken from deltawerken.com

When the water level is 2.75 metres above sea level an emergency team is physically present at the storm surge barrier. This team continuously checks the sea water level and closes the barrier in the control room (see Figure 1.3) if it is expected that the sea water level will exceed 3.00 metres above sea level. It may happen that the water level goes down even after it gets to 2.99 metres. In such cases it is forbidden to close the barrier because of the damage to the environment. Hence, before closing the barrier, the emergency team has to be sure that the sea water level will really reach or exceed 3.00 metres. When the sea water level gets exactly 3.00 metres the barrier closes automatically if no action is taken. This is an unwanted situation, since during an automatic closure the emergency team has no control.


Figure 1.3: Picture of controlling panel sluices Eastern Scheldt

Figure 1.2: Measuring device water level Roompot Buiten

# Chapter 2

# Problem description

Physical models exist that can predict reasonably well when the water level will largely exceed the 3 meters. However, when the water level stays close to the 3 meters, these physical models cannot give definite answers on whether to close the sluices or not. Therefore, a statistical model needs to be developed that can give "adequate" predictions in these boundary cases. In order to develop such a model, the following research question is posed:

> How can a reliable 5 minute-ahead predictions of the sea water level at the Roompot Buiten inner harbour of the Eastern Scheldt Storm Surge Barrier during high water be produced?

Reliable predictions should address the following issues: unwanted over- and under prediction. This raises the question when these deviating predictions are unwanted.

Over-prediction (prediction that is higher that the actual level) is unwanted when amongst the predictions, there are values that are above 3.00 meters, while the actual values stay below this threshold. In that case, the model wrongly predicts to close the storm barrier. As mentioned, this causes damage to the ecological water system.

Under-prediction, on the other hand, has unwanted consequences when the predicted values stay below 3.00 meters, while there are actual values that are above 3.00 meters. In this case, the barrier needs to be closed but the predictions indicate not. This could lead to over-flooding of the Eastern Scheldt. Different parties have different priorities, which lead them to evaluate the consequences differently. For instance, the fishery and nature activists could prefer a higher risk of flooding over the ecological consequences in the aquatic world. However, for neighbours near the storm barrier, the higher risk of flooding might be much more relevant than the consequences for the aquatic world. The Dutch Ministry of Infrastructure and Water Management evaluates these consequences as equally important. Therefore, here over-prediction will be treated equally problematic as under-prediction.

Furthermore, the prediction should satisfy the following requirements:

**R1** The predictions should include the oscillations caused by the geometry of the inner harbour

**R2** The predictions should be produced by an automated, data-driven model

**R3** The method should be able to produce new, reliable predictions within a minute

In order to arrive at a reliable prediction methodology, we will study the following sub-research questions:

**Q1** How can we develop an adaptive methodology to capture the short-term oscillations of the water level?

**Q2** How can we develop an adaptive methodology to capture the long-term trend of the tide?

# Chapter 3

# Performance measures

In Chapter 2 we posit the problem description, and came up with some requirements a prediction should satisfy. Requirement R3 does not specify what "reliable" means. In this section we will discuss possible ways to quantify what "reliable" means.

First, we have to notice that we can distinguish multiple layers in the predictions. First of all, we can talk about a five minute ahead-prediction of one single fit, and develop a performance measure for that prediction. Another layer we can identify is when we consider all single fits made during one high water, and group these single first layer predictions to form one prediction. In this research paper, we will use the terminology "peak" to describe an occurrence of high water. When the water level is considered "high" is left to the context, we will not provide one definition. We can call this grouping single first layer predictions to form one prediction described above a "prediction of the second layer", and want to develop a performance measure for this prediction as well, so that we can compare various peaks with one another. Finally, we can combine multiple first layer predictions at multiple peaks. Using another performance measure, we then obtain a peak independent error, that indicates the "reliability" of the model. For each layer, we thus need to come to a performance measure to be able to make claims about the "reliability" of the model. We will discuss the possible options and choose one performance measure for each of the layers. The different layers and their statistics become more transparent if we present them in a table:

| Layer | Perf. meas. | Fitted | Unit |
|---|---|---|---|
| 1 | $\varepsilon_{indv}^{(j)}$ | 1 model without updates, 1 peak | $y_i^{(j)(k)}$, $\widehat{y}_i^{(j)(k)}$ |
| 2 | $\varepsilon_{allvar}$ | $m$ updates of the model, 1 peak | $\varepsilon_{indv}^{(j)}$ |
| 3 | $\varepsilon$ | $m = m(nrPeaks)$ updates of the model, $nrPeaks$ peaks | $\varepsilon_{allvar}$ |

Table 3.1: Different layers that we can distinguish in the predictions and their statistics

## 3.1 First layer

For the first layer, arguably the most obvious performance measure, based on the discussion in Chapter 2 on the problems with under- and over-prediction, is that we incorporate an absolute penalty. Therefore, we check for under-prediction: we check if the first layer prediction has all their values below 3 meters, while the corresponding observed measurements has at least one value above 3 meters:

$$\varepsilon_{lwrabs} = \mathbf{1}\{\forall_{i \in prdtRng}\ \widehat{y}_i < 300\}\mathbf{1}\{\exists_{i \in prdtRng}\ y_i \geq 300\} \tag{3.1}$$

where

- $\mathbf{1}\{\}$ is the indicator function
- $i$ is a dummy variable indicating which predicted or observed measurement is considered
- $prdtRng$ is the vector containing all indices of the predicting range
- $\widehat{y}_i$ is the $i^{\text{th}}$ predicted water level
- $y_i$ is the $i^{\text{th}}$ observed water level

Next, we check for over-prediction: we check if within the first layer prediction, there exists at least one value above 3 meters, while all corresponding observed measurements are below 3 meters:

$$\varepsilon_{uprabs} = \mathbf{1}\{\exists_{i \in prdtRng}\ \widehat{y}_i \geq 300\}\mathbf{1}\{\forall_{i \in prdtRng}\ y_i < 300\} \tag{3.2}$$

with the same interpretation of the variables as with the under-prediction. The resulting performance measure then is discrete: it takes value 1 if either under- or over-prediction occurs, and 0 otherwise:

$$\varepsilon_{abs}^{(j)} = \mathbf{1}\{\varepsilon_{uprabs} = 1 || \varepsilon_{lwrabs} = 1\} \tag{3.3}$$

The main problem with this absolute performance measure is that there are almost no occurrences of high water in the data-sets of Roompot Buiten. In fact, since the opening of the sluices in 1986 until October 2019, only 27 occurrences of a water level above 3 meters occurred Wikipedia (2019). While we can make adjustments to the criterion that make it applicable to the data set (for instance by lowering the threshold level of 3 meters to a more suitable one), another criterion for the first layer is more favorable. For this criterion, we consider the relative error, that is, the difference between the predicted values and the observed values. This is done for each of the measurements. To convert these individual errors in one joint error, multiple choices, based on the different mathematical norms, exist. One popular choice, based on the 2-norm, is the sum of squares:

$$\varepsilon_{indv}^{(j)} = \sqrt{\frac{1}{n} SS_{ttl}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i^{(j)} - \widehat{y_i}^{(j)})^2} \tag{3.4}$$

where

- $j$ is a dummy variable indicating which prediction or fit is considered
- $y_i^{(j)}$ is the $i^{\text{th}}$ observation of the $j^{\text{th}}$ update of the model
- $\widehat{y_i^{(j)}}$ is the $i^{th}$ predicted or fitted value of the $j^{\text{th}}$ update of the model
- $n$ is the total number of predicted or fitted values in one prediction/fit, i.e. $6 * 5 + 1 = 31$ for a default 5 minute prediction (we start counting at "zero")

This error is somewhat relaxed; there may be instances where the difference between the observed and the predicted water level is more than 1 centimeter, as long as these instances are compensated with cases where the two coincide. This also implies that the overall error can be small while high differences between predicted and actual water level occur: if one large difference would be accompanied with otherwise exactly correct predictions, then the size of this difference can in principle be $n$. In practice, however, high water levels tend to not be isolated jumps, but rather smoothly varying and occurring in groups.

This problem vanishes when we consider the maximal difference between the predicted values and the actual values, i.e. the "$\infty$"-norm:

$$\varepsilon_{indv}^{(j)} = \max_i \left| y_i^{(j)} - \widehat{y}_i^{(j)} \right| \tag{3.5}$$

Indeed, this measure is stricter in the sense that all differences need to be within $k$ cm before $\varepsilon_{indv}^{(j)}$ is $k$ cm. The advantage is that with this measure, the point in time at which the biggest difference between the predicted and the actual water level occurs is easily determined, which possibly can give us insight on how and where to improve the prediction. The expectation is that, even with the more relaxed 2-norm, a maximal error of 1 cm will be quite ambitious. Combining this with the expectation that large differences in the prediction and actual water levels will occur in groups than to be isolated jumps, we choose the 2-norm to measure a single 5 minute prediction, i.e. a prediction on the first layer.

## 3.2 Second layer

To extend this single 5 minute prediction to multiple 5 minute predictions we need to add another "layer" of performance measures. This is needed because the time frame for which the water level is within 25 centimeters of its top typically is longer than 5 minutes. We can incorporate this extra layer of performance measures by putting a norm over the predictions, i.e. a "norm over a norm". Again, we have a choice for this second "norm". For instance, we can take the average over the performance of the individual predictions as a measure of the overall performance:

$$\varepsilon_{allavr} = \frac{1}{m}\sum_{j=1}^{m}\varepsilon_{indv}^{(j)} = \frac{1}{m}\sum_{j=1}^{m}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i^{(j)} - \widehat{y_i}^{(j)})^2} \tag{3.6}$$

Here,

- $m$ is the number of different predictions or fits considered, i.e. the time duration of the peak in minutes minus 5
- $n$ is the number of predicted or fitted values within one prediction/fit, i.e. the "length" of the predicting/fitting range
- $y_i^{(j)}$ is the $i^{th}$ observed water level in the $j^{th}$ prediction/fit
- $\widehat{y_i}^{(j)}$ is the $i^{th}$ predicted or fitted water level in the $j^{th}$ prediction/fit

The advantages and drawbacks are similar to that of the first layer: again "bad" predictions can be compensated by "good" predictions, but also, the error per prediction can become quite large. However, it is also expected that "bad" predictions will come in groups, or at least it is highly unlikely that there will be one relatively "bad" prediction accompanied with furthermore highly accurate predictions.

Also in this layer, we can use an alternative measure based on the "$\infty$"-norm:

$$\varepsilon_{allmax} = \max_{j\in\{1,\dots,m\}} \varepsilon_{indv}^{(j)} = \max_{j\in\{1,\dots,m\}} \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i^{(j)} - \widehat{y_i}^{(j)})^2} \tag{3.7}$$

It seems more natural to go for the "$\infty$"-norm, in the sense that the full prediction needs to be adequate, as it is not known beforehand when dangerous high water occurs (in Chapter 4 EDA, it can be seen that the plots confirm this: although the location of the global maximum is approximately the same, there are tops occurring after and before this maximum at multiple locations). However, again we expect that the individual differences, now meaning differences amongst the single predictions, will not be that large. The other advantage of using the average is again that the goal of predicting within 1 cm is expected to be ambitious.

## 3.3 Third layer

On top of these first two layers, we can identify a third layer, when multiple peaks are being considered. Again the two different error measures using the maximum and the average can be used, and again we choose to use the average error for the same reasons as above. All together, the error measure thus takes the form:

$$\varepsilon = \frac{1}{\ell}\sum_{k=1}^{\ell}\varepsilon_{allavr} = \frac{1}{\ell}\sum_{k=1}^{\ell}\left(\frac{1}{m}\sum_{j=1}^{m}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i^{(j)(k)} - \widehat{y_i}^{(j)(k)})^2}\right) \tag{3.8}$$

where

- $\ell$ is the amount peaks considered
- $m$ is the number of predictions in one peak
- $n = 31$ is the number of measurements in one prediction
- $y_i^{(j)(k)}$ is the $i^{th}$ observed water level in the $j^{th}$ prediction or fit of the $k^{th}$ peak
- $\hat{y}_i^{(j)(k)}$ is the $i^{th}$ fitted or predicted water level in the $j^{th}$ prediction/fit of the $k^{th}$ peak

With this error measure in hand, our goal of this paper is to develop a statistical model that give 5 minute predictions (first layer predictions) so that the error is maximally 1 cm, i.e.

$$\varepsilon \leq 1 \tag{3.9}$$

## 3.4 Conclusion

To quantify what a "reliable" prediction is, we distinguished three layers within the prediction:

- the first layer, where we considered one single five minute-ahead prediction for one peak only.
- the second layer, where we considered the predictions coming from all updates of the model, but still for one peak only.
- the third layer, where we considered the predictions coming form all updates of the model, and for multiple peaks

At each layer, a decision for a performance measure had to be made. At the first layer, we stipulated three possibilities: an absolute error, and error based on the 2-norm and an error based on the $\infty$-norm. The absolute error is impractical because there are almost no occurrences of water levels above 3 meters, and the error based on the $\infty$-norm apostatizes because it is expected to be too strict for quality that can be achieved with the model. This leaves the choice on the more relaxed 2-norm for the first layer. For the second and third layer, we proposed two options: taking the average or taking the maximum of the errors one layer lower. By the same line of reasoning as at the first layer (namely that the maximum is too strict), we chose to use the average in both cases.

In conclusion, we choose to use the following performance measure:

$$\varepsilon = \frac{1}{\ell}\sum_{k=1}^{\ell}\varepsilon_{allavr} = \frac{1}{\ell}\sum_{k=1}^{\ell}\left(\frac{1}{m}\sum_{j=1}^{m}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i^{(j)(k)} - \widehat{y_i}^{(j)(k)})^2}\right) \tag{3.10}$$

The goal of this research project is to provide 5 minute-ahead predictions (meaning first layer predictions) so that the error is maximally 1 cm, i.e.

$$\varepsilon \leq 1 \tag{3.11}$$

# Chapter 4

# Exploratory Data Analysis

In this section, we will perform an Exploratory Data Analysis. First, we numerically inspect the data set, of which significant characteristics are discussed. Some of these characteristics pose a problem, which we try to solve when we clean the data. Of main interest are the high water levels, what we call the "peaks" (as mentioned more extensively in Chapter 3). Amongst all peaks, the six most representative (which are the six highest, see the assumptions in Chapter 5) are plotted. A parabolic model will be used as first, starting model. Since the parabolic trend is assumed to approximate a sinusoidal tide, we identified inflection points within which we expect that the model is a good fit. Furthermore we separately discuss the behaviour before and after the top, as we expect (from prior papers) that the behaviour before and after the top differ quite a lot. Finally, we zoom in on the tops, to be able to describe the behaviour around the top more precisely. We then come to a conclusion on behaviour before and after the top, where we include a functional description that can be used when actually building up the model.

## 4.1 Inspecting the dataset

The Dutch Ministry of Infrastructure and Water Management has provided a data set. This data set consists of the sea level, date and time starting from 1 March 2010 up to 31 March 2011. The measurements are taken from the measuring buoy at Roompot Buiten, which is also the buoy from which the decision on closing the barrier is made (also see Chapter 1). It does, however, have its drawbacks. There are 9458 measurements that cannot be used. The greatest portion consist of unknown (NA) values, but there are also two moments (at 4 June 2010 and 10 April 2010) on which the sluices machines got cleaned, resulting in a sudden jump in the data values. The problem of the NA values is not their relative occurrence, they contribute to only 0.3 percent of the total amount of the data, but rather their bundled distribution over the data: there are large chunks of consecutive NA values. This makes it hard to replace them with appropriate values. If we remove them an unrealistic representation of the sea level is given in the plots, as a large chunk of time is removed (see Section 4.2).

## 4.2 Cleaning the data set

Now that a dataset is obtained, we need to clean it. We do not need to solve all of the problems that arise from the raw data, some can be circumvented. In this section, we will discuss what aspects of the dataset cause problems, and which of these actually get "cleaned".

As the evolution of the water level is a continuous process, we consider outliers are to be measuring points where there is a sudden jump in the water level. If these outliers are (fairly) isolated or come in "small" groups, then these outliers can be replaced by average values, interpolating between the beginning and end of the jump.

Practically, it turns out to be too complicated to replace above described type of points. We can easily identify outliers with a high value (say 600, twice the threshold for closing the sluices), but they occur seldom. In fact, there are no values above 310 and below -210, except one outlier of 9995. We can identify the other outliers, meaning the ones within the bounds of tide (approximately between 300 and -210) by comparing with predecessor and successor. We consider this not to be worthy. For some outliers we know their exact occurrence, but these are rare, making replacement unworthy.

For two occurring runs of detrended points, we know the value of the points and reason for the deviation of the trend. On 04-04-2010 and 10-04-2010, the measuring equipment was being cleaned, resulting in sudden jumps and long runs of consecutive sea levels of 300. As this includes a large part of the measurements of these days, we will not remove the measurement points of these days, but rather, we will omit these days entirely. The rest of the occurring runs of detrended points do not take a typical value or do take a NA-value and occur at a large scale of lengths, making identification difficult. Especially the treatment of the NA-values cause some

problems. We can replace the NA-values by points that respect the trend by fitting the trend of the sea level, but we will not do this as this will not lead to more adequate and realistic predictions. However, if we keep the NA values, we cannot find the peaks of the sea level, as NA values cannot be compared with the water levels. Therefore, checking whether measurements are below lower threshold level is impossible. On the other hand, if we remove the NA values, gaps in the time frames are created that lead to discontinuities in the plots and deviations in the predictions.

As a solution, we remove the NA-values before finding the peaks. Before plotting and predicting, we then reload the dataset again, including NA values.

## 4.3 Finding representative peaks

Before we can inspect the measurements graphically, we want to know which peaks are representative for high water occurrence. As we mentioned, flood-tide happens twice a day at Roompot Buiten, giving $2 \cdot 365 = 730$ potential peaks.

So which of these are representative? None of these exceed 300 centimeters, so in that sense none represents a real case of dangerous high water. We therefore need to make the assumption that the peaks with lower global maximum are representative for the dangerous high peaks. Furthermore, we assume that the higher the global maximum of the peak is, the more representative it is for high water. In the end we decided that the six highest peaks of the dataset are representative, and will work with those throughout this paper (also see assumption in Chapter 5).

## 4.4 Plots of the sea level

In this section, we show plots of the six highest peaks of the sea level at Roompot Buiten of the year 2010-2011, and discuss their characteristics such as maximal value and shape. In this manner, we hope to find general patterns present in the peaks. We will show the plots of the peaks, where we include a period of 25 minutes before the top, and 25 minutes after the top.

### 4.4.1 General form peaks

From plots of the full days (not shown in the report) we know that the tide approximately follows a sinusoid. Similar as a sinusoid, the tide also shows inflection points, i.e. points where the derivative has a global minimum or maximum. Before we show plots of these inflection points, it is useful to get an idea on how these points are found. For starters, we cannot simply move from the current measurement back in time; calculating numerical derivatives until the local maximum of the derivative is found. If the step size is too small, we can get stuck in the inflection point of a short term oscillation, which location can differ quite a lot from the location of the inflection point of the trend. If the step size is too big, we might cross the inflection point and go all the way through the first previous local minimum (ebb) to move to the next inflection point. It might very well be that in real-time predictions, peaks occur where there is no time step that solves both these problems. Instead, we identify a point $A$ in time far before the inflection point, and approach the inflection point "from the left". We do this by chopping up the time frame from that point $A$ to the current point in time into two pieces, and calculate the derivative of both. We then continue with the piece that has the largest derivative, chopping it up into two pieces and again keeping the piece that constitute to the largest derivative. We repeat this procedure a couple of times, and then save the midpoint of the piece in the last step as inflection point. Since the current moment in time can, in principle be arbitrary close to the maximum of the peak (the top), it seems to be best to pick for point $A$ the first local minimum (ebb) before the maximum of the peak. An approximate location of the local minimum suffices, the location of the found inflection point is not sensitive to small changes in the location of the local minimum (as experimenting indicates).

Plots of the estimated inflection point and local minimum (ebb) of five of the six highest peaks can be seen in Figure 4.1 up to Figure 4.5. We note that the inflection point of the first of March is missing. The reason for this is that this peak is at the beginning of the data set, so that we could not find the inflection point. The role these inflection points play in the validness of the model we will cover extensively in Section 7.2.

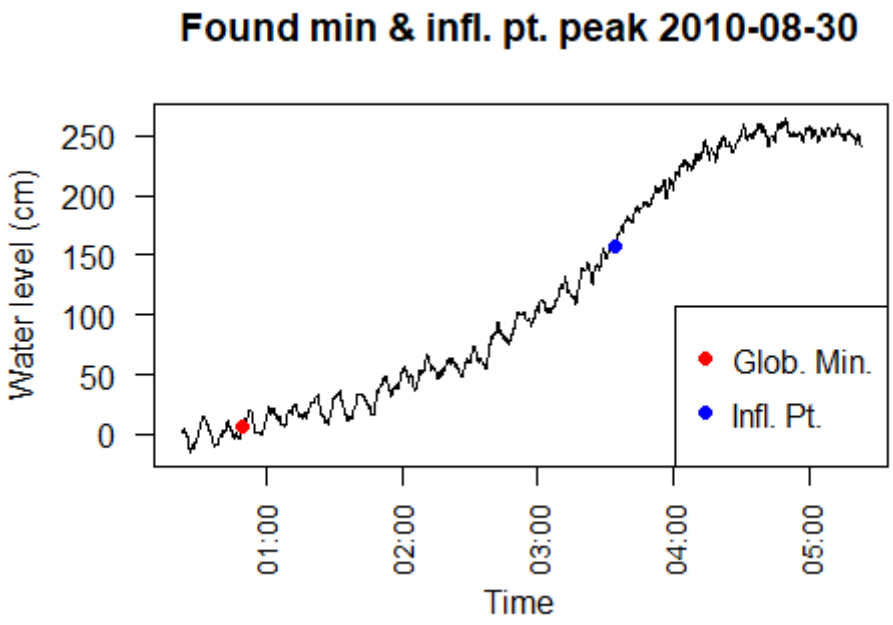


Figure 4.1: Found minimum and inflection point peak 30 August

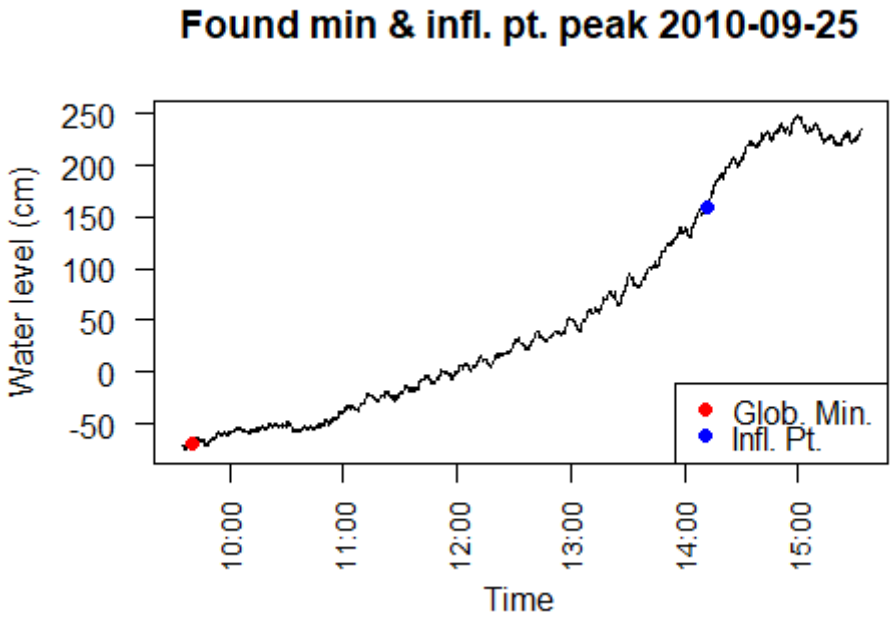


Figure 4.2: Found minimum and inflection point peak 25 September

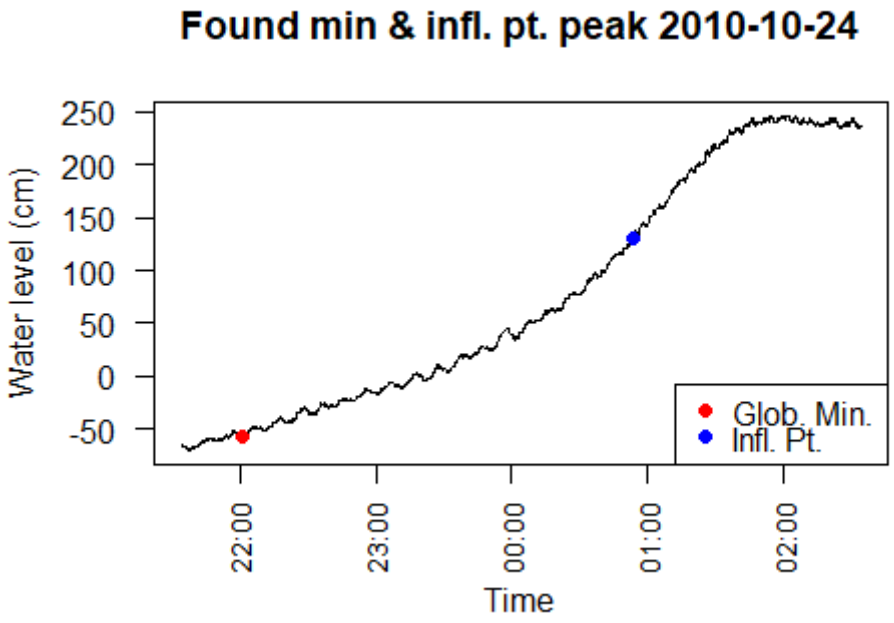


Figure 4.3: Found minimum and inflection point peak 24 October

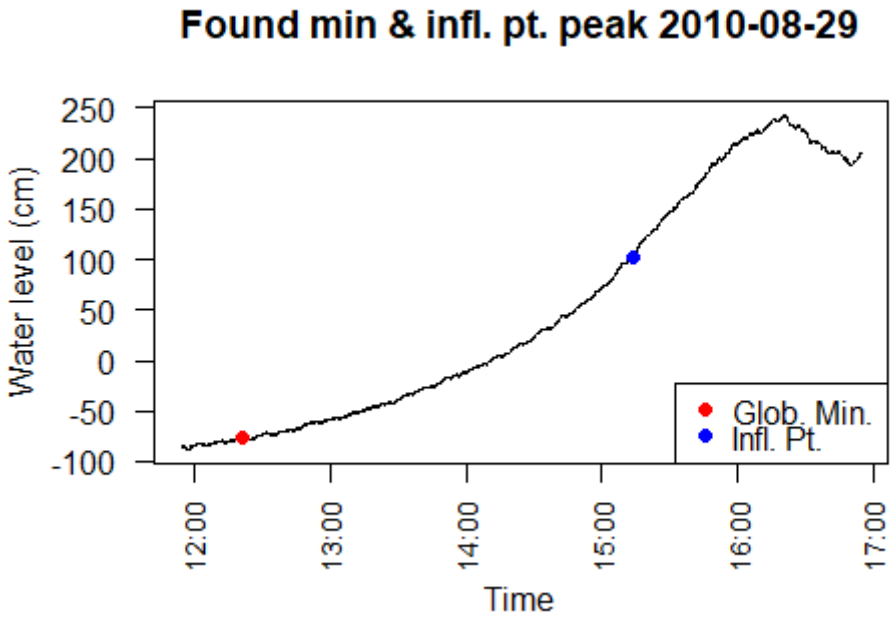


Figure 4.4: Found minimum and inflection point peak 29 August

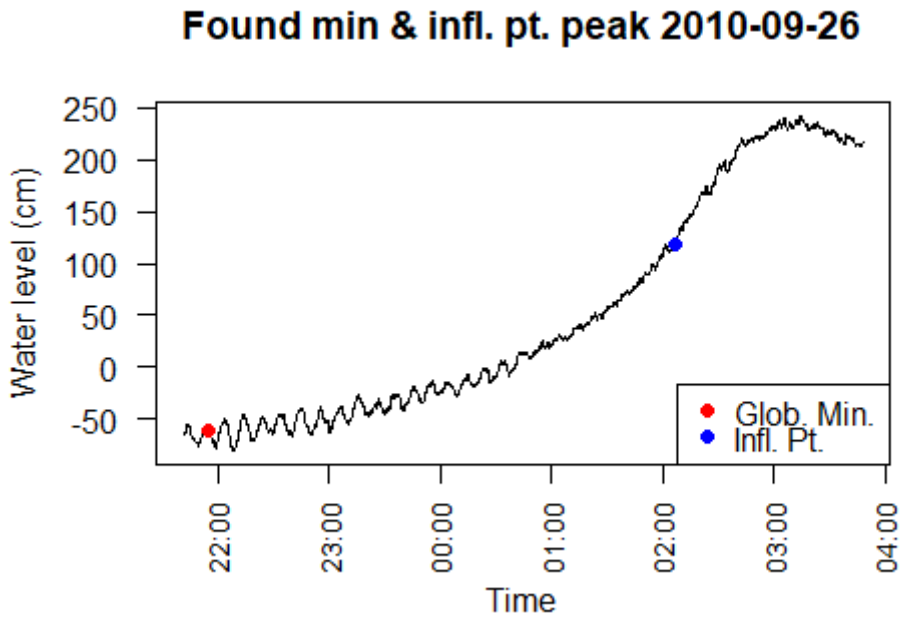


Figure 4.5: Found minimum and inflection point peak 26 September

After these inflection points, most peaks indeed follow a parabolic trend, except the fifth one, which almost follows a monotone increase. The maximum water level reached is quite similar for all the six peaks, it ranges from approximately 240 to 260. The amplitude of the short-term oscillations differs over the different peaks. The first highest peak, on 30 August, possesses the highest amplitude. On the first of March 2010, the oscillations were almost absent.

### 4.4.2 Behaviour before the top

In this section, our goal is to describe the behaviour before the top for each peak separately. We expect the trend to follow a parabola, and the short-term oscillations to be represented by sinusoids.

**First highest peak: 2010-08-30**

This peak begins with rapid, parabolic increase. In the beginning, the short-term oscillations are less "present", i.e. have smaller amplitude. The magnitude of the amplitude increases continuously. The overall trend largely represents a parabola.

**Second highest peak: 2010-09-25**

This peak has the same rapid increase to top as the peak on August 30, approximately with the same slope. Now, parabolic trend is less respected, the top is a bit higher than the parabolic shape suggests. Also, short-term oscillations seem to have larger period.

**Third highest peak: 2010-03-01**

The rapid increase that is also present in this peak continues for a longer time than at the peaks on August 30 and September 25. The slope decreases a little as the top is approached, but there is an sudden strong bend close to the top. This bend makes fitting a parabola less reasonable, although furthermore, a parabolic trend would work. Furthermore, there are low amplitude short-term oscillations, almost absent.

**Fourth highest peak: 2010-10-24**

Behaviour and shape before top are extremely comparable to the peak on March 1; this peak also has rapid increase that continues far towards the top. The peak of 24 October seem to be a zoomed in version of the peak of March 1, can hardly spot the differences. The slope decreases later as top is approached than the peak on March 1, but does so more smoothly. Also, the average slope at October 24 is lower than at March 1: here, only 15 centimeter is added to the water level while in the same time, the water level on March 1 increases by 35 centimeters. The amplitude of the oscillations is intermediate.

**Fifth highest peak: 2010-08-29**

The rapid increase that is also present in this peak gets slowed down a bit as top is approached, but does not continue to slow down all the way to the top. Almost monotonic increase, worst overall resemblance of a parabola. Top seem to appear too late, increase is already almost zero when suddenly, a second bump with the top (global maximum) occurs. In this bump, slope increases again to little local top, then a short second

increase to the global top. These two tops lie relatively close to each other. Oscillations are small sized at this peak.

**Sixth highest peak: 2010-09-26**

The shape resembles that of the peaks on August 30 and September 25, most notably September 25. First a rapid increase with some small intermediate intervals where there are one or two short-term oscillations. Then in the third intermediate interval the slope decreases a bit. From that point on wards, the parabolic shape is resembled towards the top. There are intermediate sized oscillations before the global top, creating a local top before the global top. The time between this two tops is longer than with the peak on 29 August. Furthermore, oscillations appear with the same magnitude of the amplitude throughout most of the peak, except close to the to top, where it increases as the top is approached. the magnitude of the amplitude is intermediate compared to the other peaks.

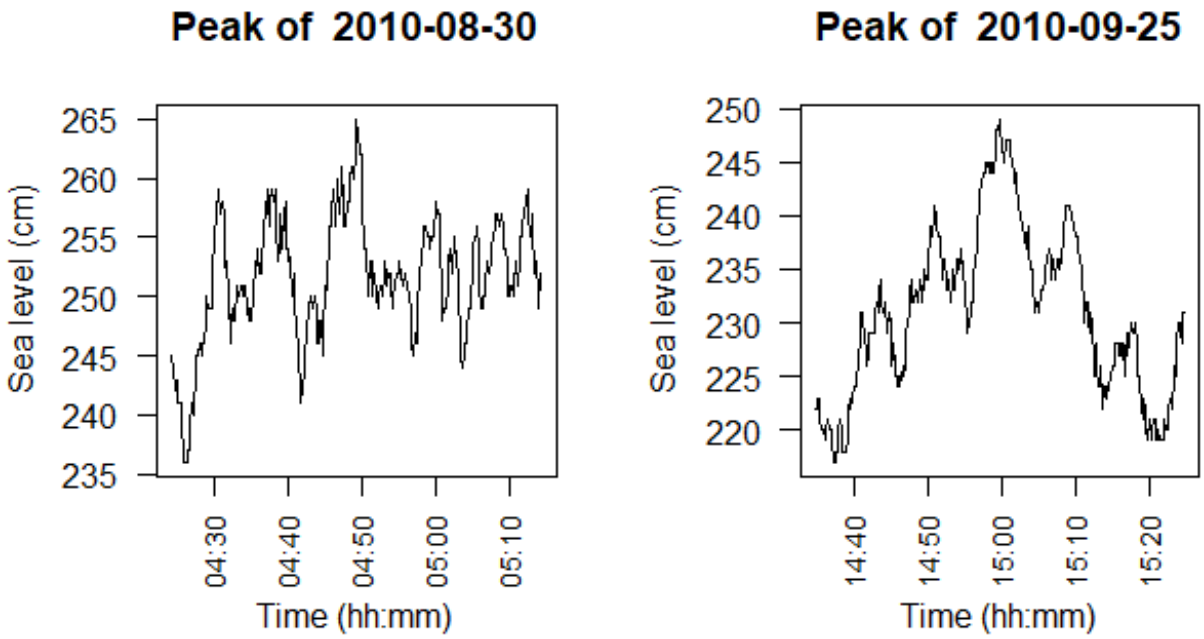


Figure 4.6: Behaviour close to the top of peaks on 30 August and 25 September

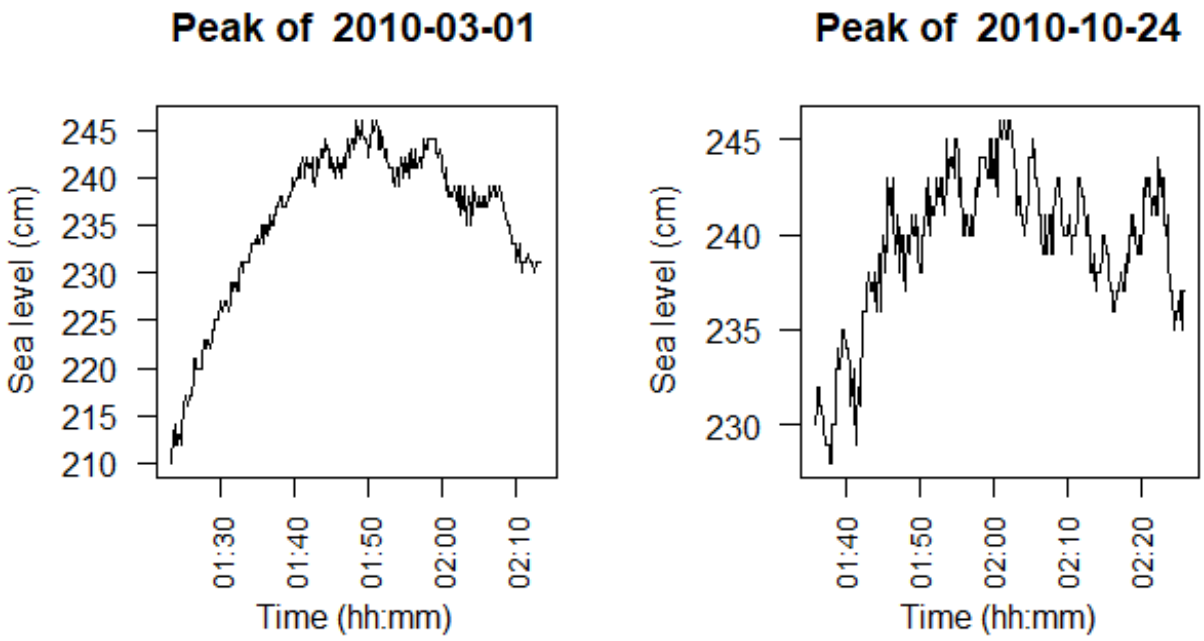


Figure 4.7: Behaviour close to the top of peaks on 1 March and 24 October

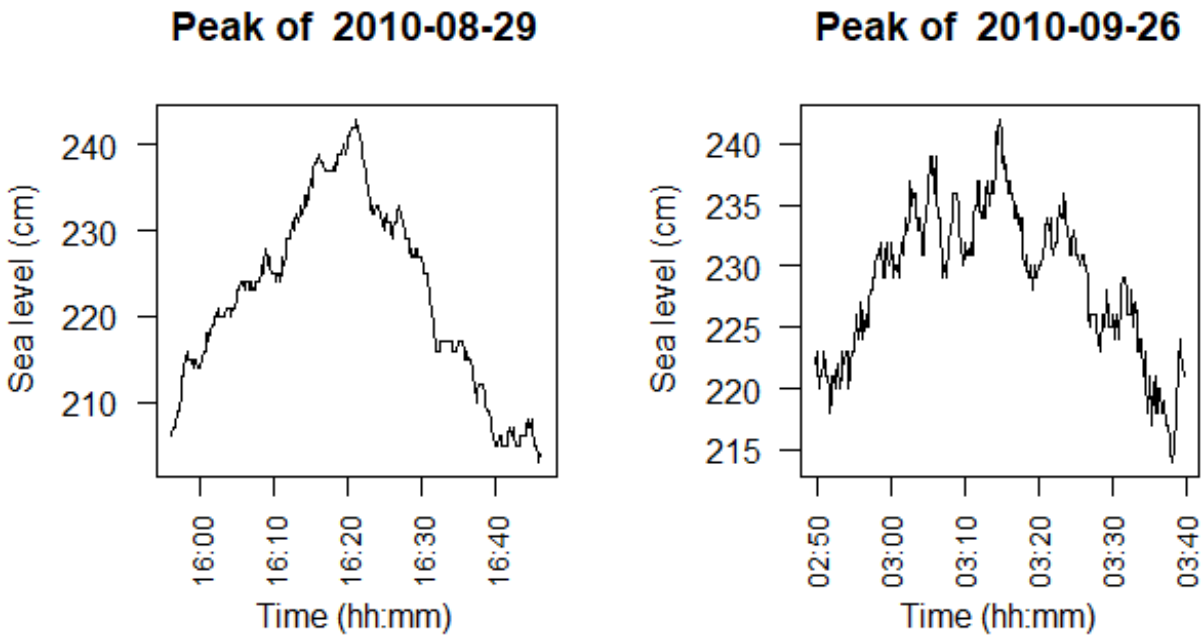


Figure 4.8: Behaviour close to the top of peaks on 29 August and 26 September

### 4.4.3 Behaviour after the top

In this section, our goal is to describe the behaviour after the top for each peak separately. We expect that it will be harder to deduce the overall behaviour of the trend and short-term oscillations, we expect them to differ extensively over the different peaks. Nevertheless, we might be able to come to two or three different forms for the trend. We expect that short-term oscillations are still well modelled by sinusoids.

**First highest peak: 2010-08-30**

After the global top there is a sudden drop in the water level, followed by low amplitude short-term os-

cillations. The amplitude increases again after a small amount of time, but the trend seem to be constant, monotonic here. It stays monotonic, but there is an instantaneous change in the slope, a nod, after which it decreases monotonously.

**Behaviour trend peak August** 30

After top trend stays constant for a while, then there is monotonic decrease.

**Second highest peak: 2010-09-25**

Directly after the top there is a small local maximum. This top seem to mainly be created by oscillations with high amplitude, which after the global maximum keep occurring with varying amplitudes. The trend seem to decrease monotonously, although there also seem to be some oscillating behaviour present in the trend.

**Behaviour peak September** 25

After top the trend follows a slightly diagonally rotated sinusoid.

**Third highest peak: 2010-03-01**

This peak shows much lower amplitudes of the short-term oscillations. There again is a parabolic shape for the trend, although it has less curvature. This parabolic shape is resembled for approximately 30 minutes, after which the trend changes to monotonic decrease. At this point, the water height forms no threat anymore. Combining this with the behaviour for the top, the trend seem to resemble a normal distribution with a skewed top. The short-term oscillations have relatively (relative to the first two peaks) small amplitudes. The highest amplitudes appear close to the top, moving away from the top in positive time direction this amplitude decreases.

**Behaviour trend peak March** 1

The trend seem to follow a normal distribution with a skewed top.

**Fourth highest peak: 2010-10-24**

This peak has a long time frame after the top where it monotonically decreases with a really small slope. As a result, the period where the sea level follows the parabola is relatively short. The oscillations have average-sized amplitudes apart from to places where the some of the oscillations seem to resonate with one another: approximately halfway the constant small decrease.

**Behaviour trend peak October** 24

After the top trend decreases monotonically.

**Fifth highest peak: 2010-08-29**

Different from the other peaks, this peak does not seem to have a parabolic shape, rather it seem to monotonically decrease with a large slope, like the graph of an absolute value. The decrease stays monotonic for a long time, after which the trend smoothly changes to being constant for a while. At this point, the water height forms no threat anymore.

The oscillations have average sized amplitudes (compared to the other peaks), although there is one occurrence of a quite large amplitude. This occurs after some oscillations with small amplitudes, possibly indicating resonance.

**Behaviour trend peak August** 29

After the top trend seem to follow a skewed normal distribution, that is, a normal distribution with top asymmetrical positioned.

**Sixth highest peak: 2010-09-26**

In this peak there seem to be less differences in size of the amplitude of the short-term oscillations. There are short periods of small amplitudes followed by relatively (relative in this peak) high amplitudes that possibly indicate resonance, but overall the amplitude seems to be the largest around the top after which it slowly decreases up to the end of the peak. A parabolic decrease seem to also be present in this peak, with approximately the same curvature as before the top.

**Behaviour trend peak September** 26

After the top, trend seem to follow a slightly diagonally rotated sinusoid.

### 4.4.4 Zoomed in six highest peaks Roompot Buiten 2010-2011

To get a better feel of the precise shape and the behaviour very nearly around the top, we zoom in on the tops, plotting the three minutes before and three minutes after the top.

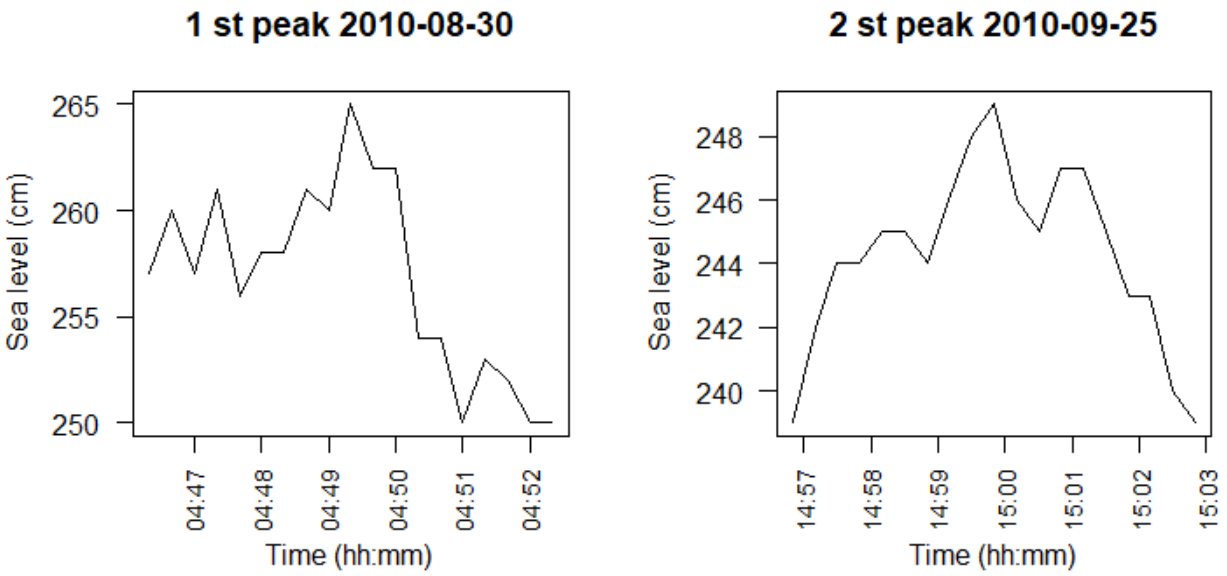


Figure 4.9: Tops of first occurring peaks 3 August and second peak 25 September

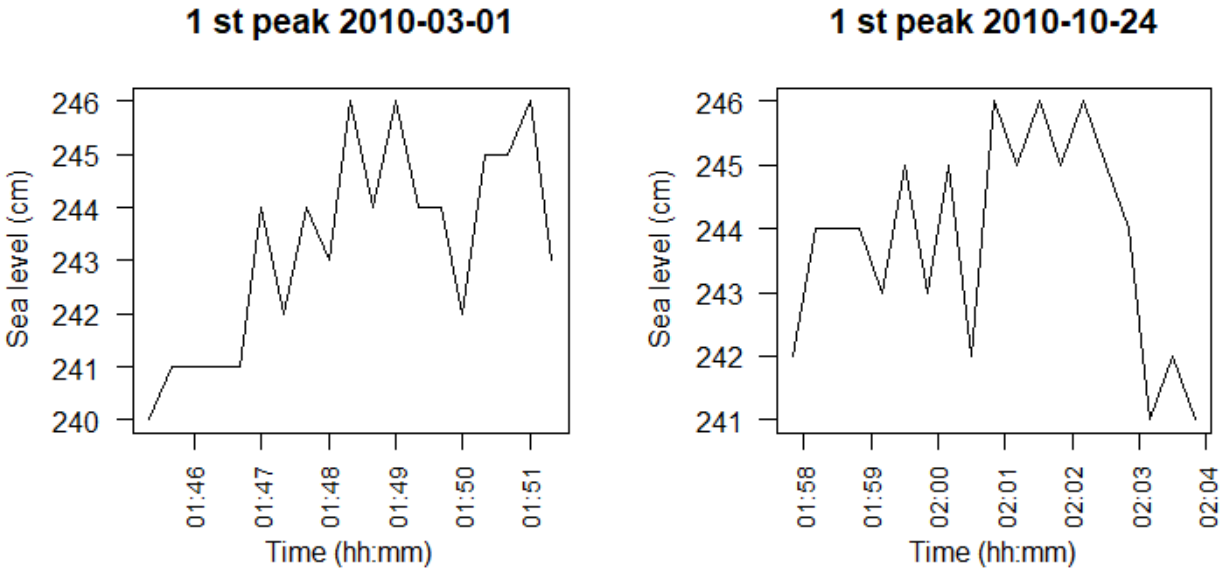


Figure 4.10: Tops of first occurring peaks 1 March and 24 October

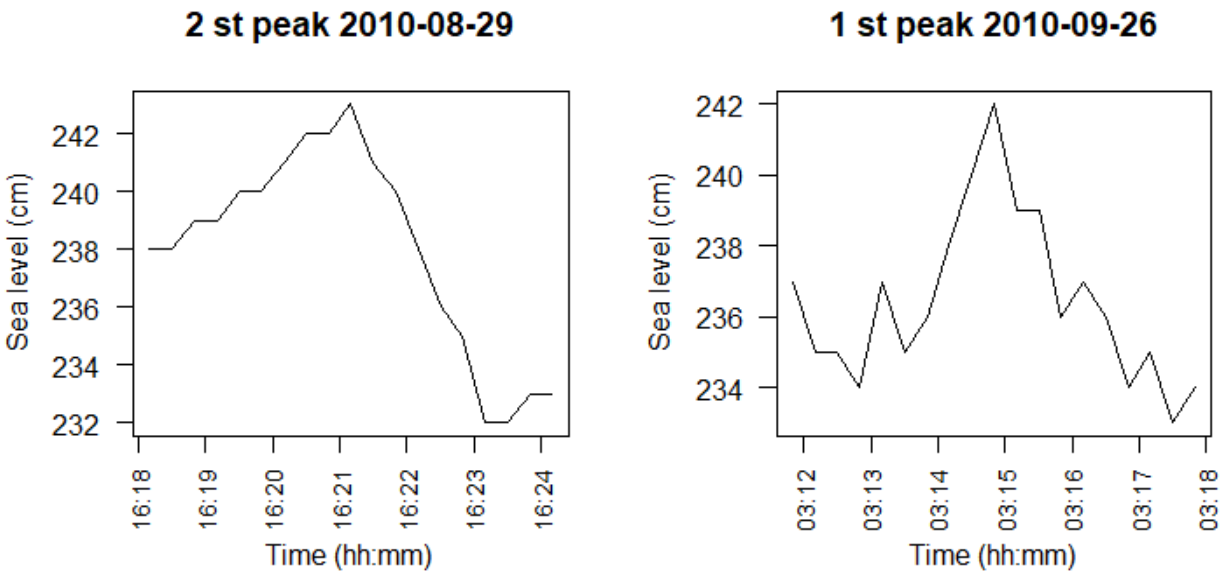


Figure 4.11: Tops of second occurring peak 29 August and first peak 26 September

Before discussing the behaviour visible on the plots, we first discuss the more global behaviour, so also the behaviour outside the time frame of the plots. The peak on 29 August behaves almost as a parabola, almost with monotone increasing and decreasing slope. There are, however, also instances of a 'pre-top' and a 'post-top', appearing separately as well as together. An extreme example is the first peak at 1 March, where there is a small 'pre-top' as well as a small 'post-top', both appearing quite far away from the top. Also the peak during second tide at 1 March behaves interestingly, having not one unique top but rather a bunch of sea levels of approximately the same level.

Next, we move on to what can be extracted from the plots. Most peaks include some local maxima, before the global maximum (meaning the top) is encountered. The number of these local maxima differs. The peak of 3 August, 25 September, 29 August and 26 September all have one top, 1 March and 24 October have three. The behaviour of the peaks with one top deviate quite a lot: the second peak on 29 August shows almost no perturbations, while in the other peaks with one global maximum some oscillating behaviour is present, sometimes accompanied with small local maxima.

## 4.5 Conclusion behaviour before and after the top

**Conclusion behaviour before the top**
All peaks show parabolic behaviour very close to the top (within 5 minutes). For most peaks, this parabolic behaviour starts quite earlier, at the "inflection points" of the water level. For those peaks, the water height is indeed well modelled by a parabola with sinusoids for the oscillations. There are, however, also instances where the parabolic behaviour is, against the expectations, not directly present after the inflection points. This is for instance the case with the peak on 29 August, that instead follows monotonic increase towards the top. The steepness of these parabolas, i.e. the focal length differs a lot for the different peaks. From the plots, it also seems as if there are large differences in the amplitude of the oscillations. In practice, there are differences (most notably between peak on August 30 and the other peaks), but they are smaller then seen at first sight. This illusion is caused by the differences in scale of the water level between different peaks. The amplitude in general seems to increase when the top is approached.
**Conclusion behaviour after the top**
The trends behaviour after the top differs quite per peak. Behaviours encountered are: monotonic decrease with a turning point, monotonic decrease with a sinusoid, monotonic decrease and Erlang distribution shaped curve. As the monotonic decrease with a sinusoid and a curve that resembles the Erlang distribution occur most often, we will try to model those. The cause of a relative high post-top local minimum close to the global minimum is the combination of (still) a high tide and a high maximum in the short-term oscillations. The behaviour of the short-term oscillations after the peak is also more difficult to describe, most notably the amplitude. Since the parameters of the sinusoid are re-evaluated every minute (when the model is refitted), we conclude that the short-term oscillations can be modelled by a sinusoid.

Finally, in Figure 4.12 and Figure 4.13 we graphically show the two suggested forms for the trend, the parabola and Erlang distribution.

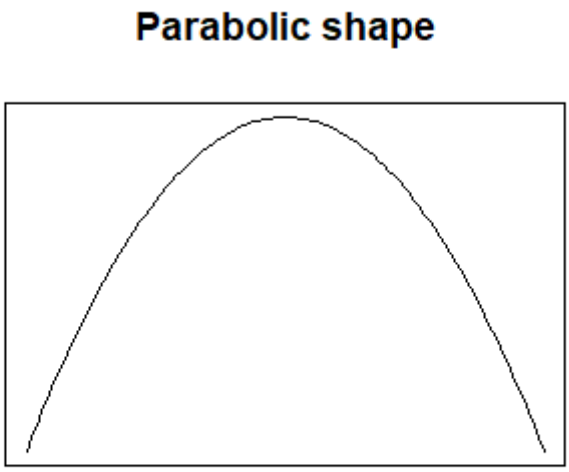


Figure 4.12: One suggested form of trend: a parabola

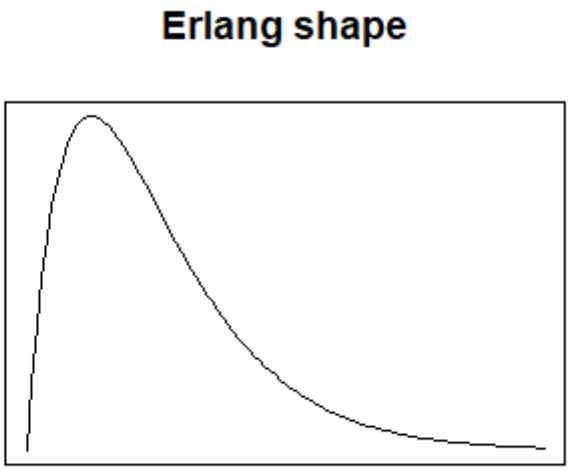


Figure 4.13: One suggested form of trend: Erlang distribution with $k = 2$ and $\lambda = 200$

# Chapter 5

# Model building

In this chapter, we will build three models to model the behaviour of the water level during high water. The first model has the same form as presented in Dassen (2016) and Bolsius (2018) and mainly serves as benchmark. The other two models arise from the systematic behaviour we have found in Section 4.4. For the independent variable time $t_i$ there are multiple choices available, dependent on which origin is picked. In Section 5.1 we discuss these options, and choose an origin for static fitting (see Chapter 6) and for dynamic fitting (see Chapter 7). In Section 5.2 we mention the assumptions necessary for the models. In Section 5.3 up to and including Section 5.5 we present the three models.

## 5.1 Choosing an origin for the time variable

When building up models, the question naturally arises what origin we should pick for the independent variable, which is the time. Arguably the most convenient choice is to take the top as origin, i.e. define

$$t_i = \text{time difference relative to the top in seconds} \tag{5.1}$$

This is convenient as for a lot of standard functions it is convenient to have the global maximum (the top) as origin.

This choice of origin is used in the static fitting. It is possible to make this choice because in static fitting, the top of the peak is known. In dynamical fitting, this is unknown (beforehand). Hence, another origin is needed. One possibility is to make use of the day time and use the beginning of the day as origin. The problem that occurs here is that then $t_i$ can take quite large values (up to 86400), especially when the peak considered occurs relatively late on the day. Moreover, the first model consists of an $t_i^2$ term, so that the corresponding coefficient has to be very small to compensate. This makes the model very sensitive to tiny changes. Another problem is that peaks occurring at two days have a jump in time.

Another possibility to overcome the unknown location of the top of the peak in the dynamic fitting is to make estimates of the location the top, and use those estimates as origin. In Section 7.3 various ways to obtain the location of the top are discussed. The main problem with this approach is that the estimate of the top is re-evaluated with a new update of the fitted model, and that these estimates differ quite a lot over all re-evaluations. As a result, two different points in time might get the same value of $t_i$, as the difference of their origins coincides with the time difference between the two points.

As an alternative, we take the beginning point of the fit range as origin in the dynamical fitting.

## 5.2 General assumptions

There are two main assumptions that all models make. These are:

- The water level can be fully explained by the sources that take time as variable. Put differently, the only independent variable is time
- The peaks found in the data set of March 2010 up to March 2011, in particular the six highest peaks with water level ranging from 2.43 to 2.65 meters, are representative for peaks with dangerous high water, that is, peaks with top higher than 3 meters

## 5.3 First model

The model we propose is a local model: each peak is modelled separately, independent of the other peaks. The tide is modelled as a sinusoid, but is thus fitted for each peak separately. At some point, the general form of

the peak deviates form the sinusoid. We locally perform a Taylor expansion around the top of the peak. Based on the discussed behaviour of the peaks in Section 4.4, our approach is to Taylor expand up to and including the second order term, modelling the tide as a parabola. In a formula:

$$\beta_0 + \beta_1 \cdot t_i + \beta_2 \cdot t_i^2 \tag{5.2}$$

By rewriting the expression, we can find an interpretation of the parameters:

$$\beta_0 + \beta_1 t_i + \beta_2 t_i^2 = \alpha_2 (t_i + \alpha_1)^2 + \alpha_0 \tag{5.3}$$

where

- $\beta_2 = \alpha_2$
- $\beta_1 = 2\alpha_0\alpha_1$
- $\beta_0 = \alpha_2\alpha_1^2 + \alpha_0$

The parameters $\alpha_0$ up to $\alpha_2$ have the following interpretation:

- $\alpha_2$ is the steepness of the parabola
- $\alpha_1$ is the horizontal coordinate of the top of the parabola
- $\alpha_0$ is the vertical coordinate of the top of the parabola

Due to the geometry of the harbor at Roompot Buiten, waves with a certain frequency start to resonate with each other, creating short-term oscillations. Following the approach of Dassen (2016) and Bolsius (2018), we model these oscillations as

$$a \sin(\omega t_i + \rho) \tag{5.4}$$

Here,

- $a$ is the amplitude of the oscillation
- $\omega$ is the frequency of the oscillation
- $\rho$ is the phase of the oscillation

Based on the prior analysis performed by the Dutch Ministry of Infrastructure and Water Management we assume that there are three main oscillations. Together, we can model these oscillations by

$$a_1 \cdot \sin(\omega_1 \cdot t_i + \rho_1) + a_2 \cdot \sin(\omega_2 \cdot t_i + \rho_2) + a_3 \cdot \sin(\omega_3 \cdot t_i + \rho_3) \tag{5.5}$$

Combining this with the tide, we come to the following model:

$$Y_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + a_1 \sin(\omega_1 t_i + \rho_1) + a_2 \sin(\omega_2 t_i + \rho_2) + a_3 \sin(\omega_3 t_i + \rho_3) + \varepsilon_i \tag{5.6}$$

## 5.4 Second model

The first model is a trend-symmetrical model with respect to the top: assuming the top of the parabola to lie close to the actual top, the behaviour after the top is modelled similarly as the behaviour before the top. In practice, however, the water level is not symmetrical: for starters, after the top the water level decreases slower than that it increases before the top. In Section 4.4, we have seen that some of the peaks show behaviour of the trend that resembles the shape of the density of an Erlang distribution. In this section, we will express the form of the Erlang distribution in its formula, and show the model that arises from this form.

### 5.4.1 Form of the model

As mentioned, in this model, we model the trend by the density of an Erlang distribution. The general formula for the probability density function of the Erlang distribution is given by:

$$f(x, k, \lambda) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!} \text{for } x, \lambda \geq 0, k \in \mathbf{N}_+ \tag{5.7}$$

The parameter $k$ is called the shape parameter and the parameter $\lambda$ the rate parameter. Indeed, by altering $k$ we obtain a different shape. After some experimenting with different values (see Figure 5.1) we find that $k = 2$

resembles the shape of the trend in the peaks the best. The parameter $\lambda$ alters the absolute location of the top, but not the shape. We let this parameter vary with the moment of the fit, so that the value will be optimized for that fit. This reduces the expression for the Erlang distribution to

$$\lambda^2 x e^{-\lambda x} \tag{5.8}$$

The shape of the Erlang distribution is meant to model the full range of the peak, so the behaviour before the top as well as after the top. This means that the Erlang distribution takes over the role of the parabola as trend. Also, to make the Erlang distribution fit the observed water level, we have to translate it. This includes a shift and a scale. The shift takes into account the relative position of the fit range with respect to the peak. The interpretation of the scale is obvious, it scales the distribution to overlap the observed water level. Written symbolically, the translation takes the following form:

$$t_i \longrightarrow \frac{t_i + A}{B} \tag{5.9}$$

where

- $A$ is the shift parameter
- $B$ is the scale parameter

In this form, the actual shift of the Erlang distribution is dependent on both $A$ and $B$. It would be more convenient if we could express the transformation in a form in which the actual shift is only dependent on the shift parameter $A$. More importantly, this other form would also be more convenient when fitting the model. The reason is that both the shift and scale parameter will appear non-linearly in the model, as they will also appear in the exponent. Thus we require starting values for them. The range of actual possible shifts is bounded: negative shifts are not allowed as the Erlang distribution is not defined for negative values and there also is a maximal positive shift that still leads to adequate fits. If there is a one-to-one relation between the shift parameter and the actual shift, than we can supply a fixed range of starting values for the shift parameter. The transition from the one form into the other is rather easy, we simply pull the shift parameter outside the fraction:

$$\frac{t_i + A}{B} = \frac{t_i}{B} + \tilde{A} \tag{5.10}$$

With this rewriting, the actual used translation of the Erlang distribution becomes:

$$t_i \longrightarrow \frac{t_i}{B} + \tilde{A} \tag{5.11}$$

Applying this translation to the expression of the Erlang distribution, we obtain:

$$\lambda^2 \left( \frac{t_i}{B} + \tilde{A} \right) e^{-\lambda\left(\frac{t_i}{B} + \tilde{A}\right)} \tag{5.12}$$

The interpretation of these parameters can be discussed shortly: the shift and scale parameters $\tilde{A}$ and $B$, coming from the translation, are already discussed. The parameter $\lambda$ is an inherit parameter from the Erlang distribution, which controls the rate.

Combining this with one extra parameter $\beta_0$, we obtain an expression for the trend of the model:

$$\beta_0 + \lambda^2 \left( \frac{t_i}{B} + \tilde{A} \right) e^{-\lambda\left(\frac{t_i}{B} + \tilde{A}\right)} \tag{5.13}$$

Next, we discuss how the parameters are used to fit this model to the observed water levels: $\beta_0$ determines the height of the initial point of the Erlang distribution. That is, without $\beta_0$, the curve of the trend would start at the end on the zero level.

The parameter $\lambda$ controls the "height" of the Erlang distribution, i.e. the distance from top to $x$-axis. A too low $\lambda$ thus causes the height to be too low. In practice this causes problems in the tail of the distribution, which is then too much stretched, so that the model deviates from the observed values too soon. On the other hand, a too high $\lambda$ causes the height to be too large, so that when shifted for a good fit, the shape becomes too narrow.

The parameter $\tilde{A}$ takes care of the horizontal positioning of the Erlang distribution. Since the density of the Erlang distribution is only defined for positive, real input, only positive shifts, i.e. shifts that move the Erlang distribution to the left are allowed. The minimal value for $\tilde{A}$ is zero, i.e. no shift. For the maximal value we take the moment the top is crossed, i.e. the maximal allowed shift is the shift where the top occurs at time is zero.

The parameter $B$ takes care of the width of the distribution: i.e. the "narrowness". We should mention that the width is also dependent on $\lambda$. But, since this parameter is used to optimize the height, we tweak the width with $B$. The boundaries of this parameter are the result of the boundaries of the other parameters, and are also found that way.

If we include the short-term oscillations in the same form as in Model 1 (Formula (5.6)), the full model becomes

$$Y_i = \beta_0 + \lambda^2 \left( \frac{t_i}{B} + \tilde{A} \right) e^{-\lambda\left(\frac{t_i}{B} + \tilde{A}\right)} + a_1 \sin(\omega_1 t_i + \rho_1) + a_2 \sin(\omega_2 t_i + \rho_2) + a_3 \sin(\omega_3 t_i + \rho_3) + \varepsilon_i \qquad (5.14)$$

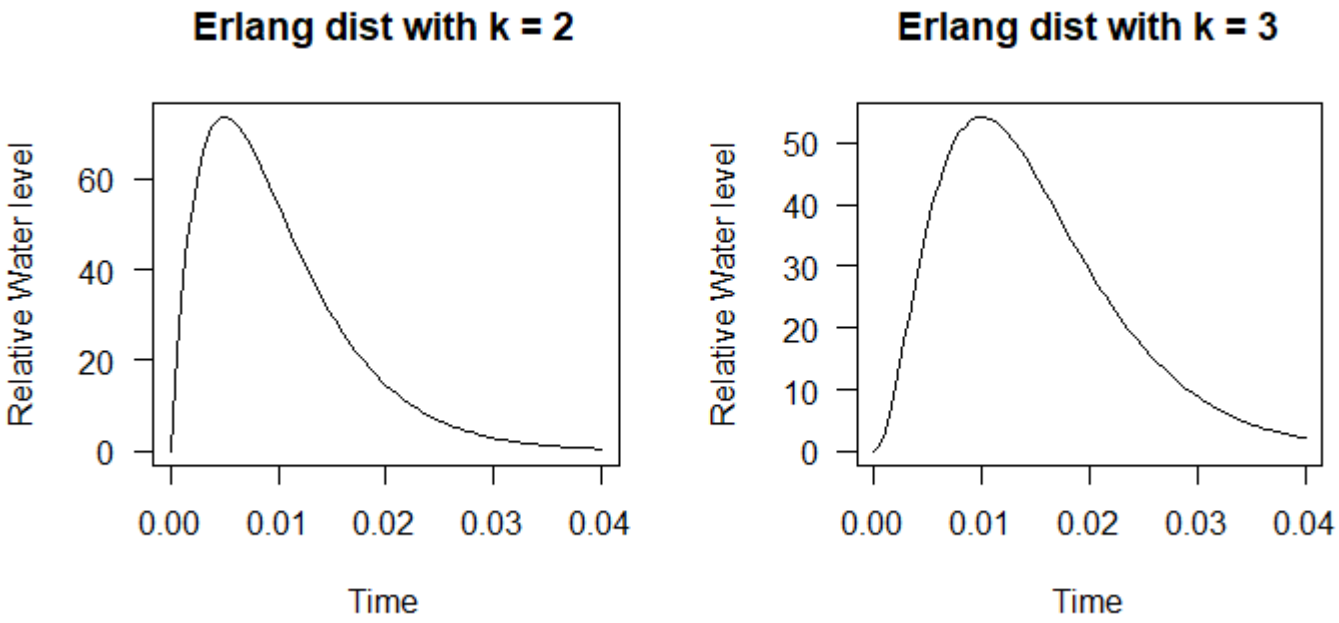


Figure 5.1: Experimenting with different values of $k$. As $k$ increases, the top becomes more symmetrical

## 5.5 Third model

### 5.5.1 Introduction

In Section 5.4 we already proposed a different model than the first, top symmetrical model to overcome the asymmetrical nature of the trend of the sea level around the top. The model uses the expression for the Erlang distribution to model the full trend, also the behaviour after the top. In Section 4.4 we have seen that the peaks of the first of March and on 29 August could be modelled by this model. Another approach to overcome the asymmetrical nature of the trend of the sea water is to maintain the parabolic trend before the top, but add another term to the model after the top, so that a two parted model is obtained. Again in Section 4.4 we saw that this is applicable to the peaks of September 25 and 26, which seem to follow a sinusoid plus a linear term after the top. In this section, we will express the form of the sinusoid plus the linear term in its formula, and show the model that arise from this form.

### 5.5.2 Form of the model

We have to notice that the two forms apply to a different time span: the sinusoid plus linear term describes the trend solely after the top. To describe the full peak including the behaviour before the top, we have to combine this with the parabolic behaviour before the top by using the indicator function. In a formula, this form of the trend can be described by:

$$(\beta_0 + \beta_1 x_i + \beta_2 x_i^2) \cdot \mathbf{1}\{x_i \leq L\} + (\gamma_0(x_i - \gamma_1) + \gamma_2 \sin(\gamma_3(x_i - \gamma_1)) \cdot \mathbf{1}\{x_i > L\} \qquad (5.15)$$

Indeed, $L$ serves as the connection point between the two different forms of the trend. The short-term oscillations are still modelled as in the first model. Together with the short-term oscillations, this model can be expressed

as:

$$Y_i = (\beta_0 + \beta_1 t_i + \beta_2 t_i^2) \cdot \mathbf{1}\{t_i \leq L\} + (\gamma_0(t_i - \gamma_1) + \gamma_2 \sin(\gamma_3(t_i - \gamma_1)) \cdot \mathbf{1}\{t_i > L\} + a_1 \sin(\omega_1 t_i + \rho_1) + a_2 \sin(\omega_2 t_i + \rho_2) + a_3 \sin(\omega_3 t_i + \rho_3) + \varepsilon_i \tag{5.16}$$

We can find an expression of $L$ by using boundary conditions on the two expressions. To let the two expressions smoothly transfer into one another, we require that the two formulas, as well as their derivatives, must be equal in the boundary point $L$. This gives us the following two equations:

$$\beta_0 + \beta_1 L + \beta_2 L^2 = \gamma_0(L - \gamma_1) + \gamma_2 \sin(\gamma_3(L - \gamma_1)) \tag{5.17}$$

$$\beta_1 + 2\beta_2 L = \gamma_0 + \gamma_3\gamma_2 \cos(\gamma_3(L - \gamma_1)) \tag{5.18}$$

This system of equations can be reduced as follows. First, the equations need to be rewritten in terms of sines and cosines, respectively. Then both equations get squared and added, so that we can use Pythagorean's theorem to reduce the sine and cosine. In that manner, we are left with one equation:

$$(\gamma_3\gamma_2)^2 = (\gamma_3\beta_0 + \gamma_3\gamma_0\gamma_1 + (\beta_1\gamma_3 - \gamma_0\gamma_3)L + \gamma_3\beta_2 L^2)^2 + (\beta_1 + 2\beta_2 L - \gamma_0)^2 \tag{5.19}$$

We can solve this equation with respect to $L$ with the help of Mathematica. If we notice we need positive solutions of $L$ (moment of time of the transition to the second form is positive) and notice that we need the positive root of the sine and cosine term (the ones we had before squaring the equations), we obtain a unique solution for $L$:

$$L = \frac{-\beta_1 + \gamma_0 + \sqrt{\beta_1^2 - 4\beta_0\beta_2 - 2\beta_1\gamma_0 + \gamma_0^2 - 4\beta_2\gamma_0\gamma_1 - \frac{8\beta_2^2}{\gamma_3} + \frac{4\beta_2\sqrt{4\beta_2^2 - (\beta_1\gamma_3)^2 + 4\beta_0\beta_2\gamma_3^2 + 2\beta_1\gamma_0\gamma_3^2 - (\gamma_0\gamma_3)^2 + 4\beta_2\gamma_0\gamma_1\gamma_3^2 + \gamma_2^2\gamma_3^4}}{\gamma_3^2}}}{2\beta_2} \tag{5.20}$$

### 5.5.3 Conclusion

We have seen three different models to model the water level during high water. The origin for the independent variable, the time, is taken to the location of the top for static fitting. For dynamical fitting, this raises a problem, as the location of the top is not known. Alternatively, here the beginning point of the fit range is taken as the origin. There are two main assumptions that all these models make:

- The water level can be fully explained by the sources that take time as variable. Put differently, the only independent variable is time
- The peaks found in the data set of March 2010 up to March 2011, in particular the six highest peaks with water level ranging from 2.43 to 2.65 meters, are representative for peaks with dangerous high water, that is, peaks with top higher than 3 meters

The models differ in their approach to model the trend (the long-term tide). In all models, the short-term oscillations are modelled by

$$a \sin(\omega t_i + \rho) \tag{5.21}$$

The first model models the tide as a parabola. The model can be expressed as

$$Y_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + a_1 \sin(\omega_1 t_i + \rho_1) + a_2 \sin(\omega_2 t_i + \rho_2) + a_3 \sin(\omega_3 t_i + \rho_3) + \varepsilon_i \tag{5.22}$$

The main disadvantage of the first model is that it models the trend symmetrically with respect to the top, while we have seen in Chapter 4 that the six peaks with the highest maximal water level all exhibit an asymmetrical trend: the trend for the top differs quite in shape from the trend after the top. To take into account this asymmetrical behaviour, we propose a different model. This model models the trend by means of the density of the Erlang distribution. After some experimenting with different values of $k$ we came to the conclusion that the shape of $k = 2$ resembles the shape of the peaks the most. The full form of the model is

$$Y_i = \beta_0 + \lambda^2 \left( \frac{t_i}{B} + \tilde{A} \right) e^{-\lambda\left(\frac{t_i}{B} + \tilde{A}\right)} + a_1 \sin(\omega_1 t_i + \rho_1) + a_2 \sin(\omega_2 t_i + \rho_2) + a_3 \sin(\omega_3 t_i + \rho_3) + \varepsilon_i \tag{5.23}$$

Finally, we propose a different model to deal with the asymmetrical nature of the trend. This model is splitted into two parts: the familiar parabola before and some time after the top, and a combination of a linear term and a sinusoid for the remaining part after the top. The model takes the following form:

$$Y_i = (\beta_0 + \beta_1 t_i + \beta_2 t_i^2) \cdot \mathbf{1}\{t_i \leq L\} + (\gamma_0 (t_i - \gamma_1) + \gamma_2 \sin(\gamma_3 (t_i - \gamma_1))) \cdot \mathbf{1}\{t_i > L\} \tag{5.24}$$

$$+ a_1 \sin(\omega_1 t_i + \rho_1) + a_2 \sin(\omega_2 t_i + \rho_2) + a_3 \sin(\omega_3 t_i + \rho_3) + \varepsilon_i \tag{5.25}$$

With these three models we conclude that we can express the found shapes in Chapter 5 in a mathematical model.

# Chapter 6

# Static model fitting

## 6.1 Introduction

In the previous chapter, Chapter 5 Model building, we build up models that should be able to model the water level during high water. In this chapter, we will start fitting these models. We distinguish two different ways of model fitting: static model fitting and dynamic model fitting. With static model fitting we mean fitting the model once over the full peak, that is: use all measurements of the peak to fit the model once, investigate the "goodness of fit" of the model and test (the form of) the model itself. In dynamical model fitting, on the other hand, the model is fitted on a smaller fit range and updated every minute, to mimic the real-time situation. In this chapter, we will focus on the static model fitting. First, we will discuss some practicalities that need to be dealt with first before the fitting can be done. In Section 6.7 and Section 6.8, the actual static fitting is done. In these sections, we keep the notions and terms informal and un-quantified on purpose. The investigation serves merely as to get a sense on whether the modelled forms of the water level are applicable, by considering their graphs.

## 6.2 Classification parameters

In the formula of a statistical model we can distinguish three different types of parameters: those that appear linearly in the model, those that appear conditionally linear in the model and those that appear non-linearly in the model. We can make the distinction on being linear, conditionally linear or non-linear by considering the derivative with respect to the parameter. If the differentiated expression of the model with respect to parameter does not contain any parameters anymore, we say that parameter "appears linearly in the model". If the differentiated expression does still contain parameters, but not the differentiated parameter, we say the parameter "appears conditionally linear in the model" (conditioned on the parameters that remained present). If the differentiated expression still contains the differentiated parameter, we say that parameter "appears non-linearly in the model".

In model 1 (see Formula (5.6)), we can see that the three parameters in front of the second degree polynomial modelling the trend: $\beta_0$, $\beta_1$ and $\beta_2$ appear linearly in the model. For the short-term oscillations, the situation is different. Given that we know the expression inside the sinusoid, the coefficient in front of the sinusoid, the amplitude, is a linear parameter. Thus we conclude that the amplitudes $a$ appear conditionally linear in the model. The parameters inside the sinusoid, the frequency $\omega$ and the phase $\rho$, appear non-linearly in the model.

In model 2 (see Formula (5.14)), there are also three parameters in the Erlang distribution modelling the trend: the rate parameter $\lambda$ and shift parameter $\tilde{A}$ and the scale parameter $B$. All three parameters appear in the exponent, so that all three parameters appear non-linearly. The short-term oscillations are modelled in the same manner as in model 1, so those parameters appear in the same manner.

## 6.3 Reducing non-linear parameters

### 6.3.1 Rate parameter $\lambda$ in Trend Erlang distribution

As discussed when we build up the second model (see Section 5.4), the part modelling the trend consists of three parameters: the rate parameter $\lambda$, a shift parameter $\tilde{A}$ and a scale parameter $B$. All these three parameters appear non-linear in the model. In general, we can reduce the number of non-linear parameters by expressing parameters in terms of others (and possibly the independent variable), by using properties of the model. In this case, we can express the rate parameter $\lambda$ in terms of the shift parameter $\tilde{A}$ and the scale parameter $B$. This

is done using the condition that the derivative should be zero at the top: if we write

$$f(\lambda, x) := \beta_0 + \lambda^2 \left(\frac{x}{B} + \tilde{A}\right) e^{-\lambda\left(\frac{x}{B}+\tilde{A}\right)} \tag{6.1}$$

then

$$\frac{\partial f}{\partial x} = \frac{1}{B}\lambda^2 e^{-\lambda\left(\frac{x}{B}+\tilde{A}\right)} - \frac{1}{B}\lambda^3 \left(\frac{x}{B} + \tilde{A}\right) e^{-\lambda\left(\frac{x}{B}+\tilde{A}\right)} \tag{6.2}$$

$$= \frac{1}{B}\lambda^2 \left(1 - \frac{\lambda x}{B} - \lambda\tilde{A}\right) e^{-\lambda\left(\frac{x}{B}+\tilde{A}\right)} \tag{6.3}$$

Setting the derivative equal to zero, we obtain

$$1 - \frac{\lambda x}{B} - \lambda\tilde{A} = 0 \implies \lambda = \frac{1}{\frac{x}{B} + \tilde{A}} \tag{6.4}$$

Thus, an estimator for $\lambda$ is given by

$$\widehat{\lambda} = \frac{1}{\frac{x_{max}}{B} + \tilde{A}} \tag{6.5}$$

Notice that above reduction of $\lambda$ in terms of $x_{max}$, $B$ and $\tilde{A}$ can only be practically applied when $x_{max}$ is known, i.e. when the location of the top of the peak is known. For static fitting, this is indeed the case. For dynamic fitting, however, this is not the case, and we will keep the second model in the form of Formula 5.14 and also supply starting values for $\lambda$.

With this reduction, the formula of the second model turns into

$$Y_i = \beta_0 + \left(\frac{1}{\frac{t_{max}}{B} + \tilde{A}}\right)^2 \left(\frac{t_i}{B} + \tilde{A}\right) e^{-\left(\frac{1}{\frac{t_{max}}{B}+\tilde{A}}\right)\left(\frac{t_i}{B}+\tilde{A}\right)} + a_1 \sin(\omega_1 t_i + \rho_1) + a_2 \sin(\omega_2 t_i + \rho_2) + a_3 \sin(\omega_3 t_i + \rho_3) + \varepsilon_i \tag{6.6}$$

### 6.3.2 Phases $\rho$ in short-term oscillations

As mentioned, we model the short-term oscillations as sinusoids of the form $a\sin(\omega t + \rho)$. Previous proposed regression models of Dassen (2016) and Bolsius (2018) adhere to this form. A disadvantage lies in the non-linear character of the parameters: in this form, both the frequency and the phase will appear non-linear in the model and therefore need to be pre-optimized before we can fit the model. This can be undesirable, as this pre-optimization is computationally expensive, both in space and time. An alternative to this form, mentioned in the paper of Ramaekers and Michels (2019) , relies on the summation formula of the sine:

$$a_j \cdot \sin(\omega_j \cdot t + \rho_j) = a_j \cdot \sin(\omega_j \cdot t)\cos(\rho_j) + a_j \cdot \cos(\omega_j \cdot t)\sin(\rho_j) \tag{6.7}$$

By identifying

$$\alpha_j = a_j \cdot \cos(\rho_j) \tag{6.8}$$

and

$$\beta_j = a_j \cdot \sin(\rho_j) \tag{6.9}$$

the representation of the oscillation in the form of RHS Formula (6.7) is linear in parameters $\alpha_j$ and $\beta_j$:

$$a_j \cdot \sin(\omega_j t + \rho_j) = \alpha_j \cdot \sin(\omega_j t) + \beta_j \cdot \cos(\omega_j t) \tag{6.10}$$

An advantage of this form is the decrease in the amount of non-linear parameters: now only the frequencies $\omega_j$ need to be optimized numerically. The downside is that we replaced each single phase $\rho_j$ with two conditionally linear parameters $\alpha_j$ and $\beta_j$ so that the total amount of linear parameters increases, but these need no separate optimization. If we apply this rewriting to all three oscillation terms, we can reduce the total amount of non-linear parameters (for the oscillation part) by three. It should be noted that there is a non-linear relationship between parameters $\alpha_j$ and $\beta_j$ that becomes apparent if we interpret the two as functions of $\rho_j$:

$$\alpha_j^2 + \beta_j^2 = a_j^2 \cdot \cos(\rho_j)^2 + a_j^2 \cdot \sin(\rho_j)^2 = a_j^2 \tag{6.11}$$

It seems that this relationship constrains the parameter space: the extra equation reduces the degrees of freedom from two to one. However, the amplitude $a_j$ is a free parameter that can take any non-negative value so that actually, there is no constraint on $\alpha_j$ and $\beta_j$. For the first model in practise there is no substantial gain with this alternative form, the model with the oscillations in the original form already converges fast enough (within 1 minute). For the second model we postpone the rewriting to future research. In conclusion, we thus stick to the original form.

## 6.4 Determination starting values for non-linear parameters

The classification of the parameters in three types linear, conditionally linear and non-linear (see Section 6.2) becomes important when fitting the model. If there are non-linear parameters present in the model, we cannot fit the model by using the default least squares method, but rather need a non-linear least square optimization. This numerical optimization requires starting values for the non-linear parameters.

In our proposed models (see Formula 5.6 and Formula 5.14), the frequency and phase of the sinusoids modelling the short-term oscillations are non-linear parameters. Also, the shape parameter $\lambda$ and the shift and scale parameters $\tilde{A}$ and $B$ of the Erlang distribution appear non-linear in the model including the Erlang distribution. In this section, methods are described to obtain starting values.

### 6.4.1 The height $\beta_0$

It might seem counter-intuitive that the height needs starting values, as it appears linear in Formula 5.14. We need to remember, however, that in the search of the values for the parameters of the trend that lead to the optimal fit, all parameters need to be optimized at the same time. That is, if we would first optimize the rate $\lambda$, shift $\tilde{A}$ and the scale $B$ together and then afterwards vary the height $\beta_0$, it is very well possible that we do not find the optimal fit. However, if $\beta_0$ is not treated as a non-linear parameter, than this unwanted situation occurs: the parameters $\lambda$, $\tilde{A}$ and $B$ are optimized firstly together, but the parameter $\beta_0$ is optimized seperately afterwards. Alternatively, we multiply $\beta_0$ with another parameter $\beta_1$ to turn it into a non-linear parameter, and supply a range of starting values for $\beta_1$.

### 6.4.2 The scale $B$

The scale parameter in the translation changes the width as well as the height of the distribution, just like the shape parameter $\lambda$, but in a different fashion. Nevertheless, as we have used $\lambda$ for the height of the distribution, we will use $B$ for the width of the distribution. The boundaries of this parameter are the result of the boundaries of the other parameters, and are also found that way. We again supply a range of starting values.

### 6.4.3 The shift $\tilde{A}$

In Section 5.4 we discussed that instead of taking the shift parameter relative to the scale parameter by dividing it to the scale (as is common practice), we separated the shift parameter from the scale parameter. We also already mentioned that this is convenient for the supplying a starting value for $\tilde{A}$, as the allowed range is fixed, i.e. independent of the scale. The minimal value is zero, as the Erlang distribution is not defined for negative values. The maximal allowed shift takes the top as starting point. Choosing a suitable time step, we obtain a range of starting values for the shift parameter $\tilde{A}$.

### 6.4.4 The phase $\rho$

As already mentioned, in the form $a\sin(\omega t + \rho)$, the additive parameter $\rho$ has the pleasant interpretation of being the phase of the oscillation. We can use this interpretation to supply starting values. Since the phase of an oscillation always lies between 0 and $2\pi$, the range of starting values is bounded. Therefore, instead of supplying a single starting value, it is possible to supply a list of starting values that goes through the full range. The default used list splits the range $[0, 2\pi]$ into 5 parts and thus contains 5 equally distant starting values.

### 6.4.5 The frequency $\omega$

The frequencies $\omega$ of the short-term oscillations appear non-linearly in the model and thus we need estimates as starting values for the program. We will present two ways of finding these estimates: one based on a local Taylor expansion and the other on using a so-called periodogram. A periodogram, informally speaking, is a graphical tool we can use to trace down the dominant frequencies in a signal. Both methodologies require the signal to be detrended. After experimenting with both we decided that it is better to work with the periodogram.

#### Local Taylor expansion

The idea with a local Taylor expansion is that the separate tops visible in the plot of the detrended data are a small part of one particular oscillation with a certain frequency. By (locally) fitting a Taylor polynomial over these separate tops a substantial amount of frequencies can be found. We can reduce this large amount of frequencies to a small amount (in principle three) of main frequencies by grouping together the frequencies of approximately the value. The individual frequencies are obtained as follows:

The Taylor expansion takes the form

$$a\sin(\omega t+\rho) \approx a\sin(\omega t_{min}+\rho) + a\cos(\omega t_{min}+\rho)\omega(t-t_{min}) - a\sin(\omega t_{min}+\rho)\omega^2\frac{(t-t_{min})^2}{2} \tag{6.12}$$

Rewriting this in powers of $t$ we obtain

$$\begin{aligned} a\sin(\omega t+\rho) \approx\ & a\sin(\omega t_{min}+\rho) - a\cos(\omega t_{min}+\rho)\omega t_{min} - a\sin(\omega t_{min}+\rho)\omega^2\frac{t_{min}^2}{2} \\ & + \big(a\cos(\omega t_{min}+\rho)\omega + a\sin(\omega t_{min}+\rho)\omega^2\big)\cdot t \\ & - \big(a\sin(\omega t_{min}+\rho)\frac{\omega^2}{2}\big)\cdot t^2 \end{aligned}$$

The fitted model takes the form:

$$Y_i = \alpha_0 + \alpha_1 t + \alpha_2 t^2 \tag{6.13}$$

Thus, if above model is fitted, we obtain the following relations:

$$\begin{cases} \alpha_0 = a\sin(\omega t_{min}+\rho) - a\cos(\omega t_{min}+\rho)\omega t_{min} - a\sin(\omega t_{min}+\rho)\omega^2\frac{t_{min}^2}{2} \\ \alpha_1 = a\cos(\omega t_{min}+\rho)\omega + a\sin(\omega t_{min}+\rho)\omega^2 \\ \alpha_2 = a\sin(\omega t_{min}+\rho)\frac{\omega^2}{2} \end{cases}$$

By rewriting and substituting, we obtain

$$\begin{cases} \alpha_0 = \frac{\alpha_2\cdot 2}{\omega^2} - \alpha_1\cdot t_{min} + \alpha_2\cdot t_{min}^2 \\ \alpha_1 - \alpha_2\cdot 2t_{min} = a\cos(\omega\cdot t_{min}+\rho)\cdot\omega \\ \alpha_2 = a\sin(\omega t_{min}+\rho)\frac{\omega^2}{2} \end{cases}$$

By rewriting the first equation, we can find an estimate for $\omega$:

$$\omega = \sqrt{\frac{2\alpha_1}{\alpha_0 + \alpha_1\cdot t_{min} - \alpha_2\cdot t_{min}^2}} \tag{6.14}$$

**Periodogram**

This part is largely based on Section 2.4.1 of the paper of Dassen (2016). Before we formally define a periodogram, we discuss its informal notion. Informally, the periodogram is graphical tool used to trace down the dominant frequencies in a signal. The main approach is as follows. For each frequency, we calculate the length of their discrete Fourier transform and use that as a measure to compare their relative presence in the signal. The plot made shows the frequencies plotted against the length of their discrete Fourier transform. By extracting the frequencies with the three highest values in the spectrum, we can obtain estimates for the frequency. We will use these estimates directly as starting values in the model.

To formally define the periodogram, we first notice that any oscillation in form of a (co)sine can be written as a combination of a sine and a cosine. (Shumway and Stoffer (2016) Section 4.1). Let $x(t)$ denote the value of the cosine at time $t$, then

$$x(t) = a\cos(2\pi\omega t+\varphi) = a(\cos(2\pi\omega t)\cos(\varphi) - \sin(2\pi\omega t)\sin(\varphi)) = U_1\cos(2\pi\omega t) + U_2\sin(2\pi\omega t) \tag{6.15}$$

where $U_1 = A\cos(\varphi)$ and $U_2 = -A\sin(\varphi)$. It holds that $A = \sqrt{U_1^2+U_2^2}$ and $\varphi = \arctan(\frac{U_2}{U_1})$. In general, for $q$ cosines, with possibly different frequencies and amplitudes (Shumway and Stoffer (2016) Section 4.1)

$$x(t) = \sum_{k=1}^{q}[U_{2k-1}\cos(2\pi\omega_k t) + U_{2k}\sin(2\pi\omega_k t)] \tag{6.16}$$

The discrete Fourier transform ($DFT$) in R is defined by (Shumway and Stoffer (2016) Section 4.3)

$$d(\omega_j) = \sum_{t=1}^{n} x_t e^{2\pi\omega_j(1-t)} \tag{6.17}$$

where $\omega_j = \frac{j}{n}$ are called the Fourier frequencies. The periodogram is then defined (Shumway and Stoffer (2016) Section 4.3) as $I(\omega_j) = |d(\omega_j)|^2$

The more dominant the frequency $\omega_j$ is present in the signal, the higher the value of $I(\omega_j)$ (Shumway and Stoffer (2016) Section 4.3). Also frequencies close to the highly present frequency tend to have high values in the

periodogram. For this reason it can be hard to pin down the exact frequency based on the plot. To overcome this problem kernel smoothing in the periodogram can be used. The kernel estimator is given by

$$\widehat{f}_n(x) := \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-X_i}{h}\right) \tag{6.18}$$

where $K$ is a kernel function and $h$ is the bandwidth.

## 6.5 Detrending the data

### 6.5.1 Introduction

In Section 6.4.5 we presented two ways of finding starting values for the frequency. These methods require a detrended fit range to obtain estimates for the frequencies. In this section, we first develop a numerical performance measure, so that it can automatically be checked whether a detrend is considered good. Then we discuss the approach and possible issues of detrending relative to the first model. The second model follows the same approach.

### 6.5.2 Performance measures for detrending the data

Since different models have different forms for the trend, the detrending of the data is model-dependent. In general, we use the same formula for the trend in the detrending as stipulated in the model (indeed, for a parabolic modelled trend, the detrended signal is obtained by subtracting a second order polynomial from the original data). This also means that we can only locally detrend the water level, after all, the modelled form of the trend was only valid locally in the first place. To find out how well a given detrended part of the data is detrended, a first heuristic approach is that we inspect the plot of the detrended data. For well detrended data the expectation is that the detrend is centered around 0 and that the shape resembles that of a linear combination of 3 sinusoids. To further compare this we can plot the detrended data together with a linear combination of 3 sinusoids with the optimized values for the parameters.

The drawback of this approach is that it is solely heuristic, and very time consuming when more peaks are considered. To investigate this more systematically as well as formally, we need to develop performance measures for the quality of a detrend. One approach, suggested in Dassen (2016), is that we fit a polynomial model on the detrended data (in the paper a first degree model is used). If the coefficients are close enough to zero, that is, the p-values on the hypothesis

$$H_0 : \beta_i = 0 \text{ versus } H_a : \beta_i \neq 0 \tag{6.19}$$

are higher than say, $1-\alpha = 0.95$ (where $\alpha = 0.05$ is the significance level), then the detrend is considered correct.

When we experiment with this approach it turns out not to work well in practice: for low order polynomial up to third degree, the coefficients stay close to zero while graphically a trend can be seen (the detrend clearly is not centred around 0). On the other hand, polynomials from third degree on wards give significant coefficients even when the detrend graphically seems okay.

As an alternative, we can use a performance measure based on an average. The first idea is that we simply take the total average of the detrended data. In a formula, this would be

$$perfdtrnd = |\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)| \tag{6.20}$$

where

- $y_i$ is the $i^{th}$ considered water level
- $\widehat{y}_i$ is the $i^{th}$ by the trend predicted water level
- $n$ is the amount of measurements in the fit range (= 211 for a fit range of 35 minutes)

Although this average would trace down a trend with a period approximately equal to the time range of the detrended data itself, we might not be able to spot trends with smaller periods, for instance half of the total time range, as the (too) high values cancel with the (too) low values. We can overcome this issue by not only taking the total average, but also the maximum average of the both halves of the detrended data, the maximum average of all three thirds of the detrended data, etcetera. In general, we subdivide the detrended data in $k$ pieces. For every piece seperately, we calculate the average. When we have collected all averages, we calculate

the maximum over all averages. In formulas, we can obtain the performance measure as follows: First, we give an expression for the error per piece of a split:

$$errdtrnd_j^{(\ell)} = |\frac{1}{n_j}\sum_{i=1}^{n_j}\left(y_i^\ell - \widehat{y}_i^\ell\right)| \tag{6.21}$$

where

- $errdtrnd_j^{(\ell)}$ is the error of the $j^{th}$ piece of a split at split $l$
- $n_j$ is the number of measurements in the $j^{th}$ piece of a split
- $y_i^\ell$ is the $i^{th}$ observed water level in the $j^{th}$ piece of a split
- $\widehat{y}_i^\ell$ is the $i^{th}$ by the trend predicted water level in the $j^{th}$ piece of a split

Indeed, it holds that $n_1 = n_2 = ... = n_{k-1} \neq n_k$, i.e. apart from the last piece of a split that deals with the remaining observations, all other pieces are equally large. Then, we can calculate the error per split (taking the maximum over the pieces):

$$errdtrnd^{(\ell)} = \max_{j\in\{1,..,k\}} errdtrnd_j^{(l)} \tag{6.22}$$

Here:

- $errdtrnd^{(\ell)}$ is the error of split $\ell$
- $k$ is the number of pieces in a split

Finally, we can give an expression for the total error of the detrend:

$$perfdtrnd = \max_{l\in\{1,..,m\}} errdtrnd^{(l)} \tag{6.23}$$

where

- $perfdtrnd$ is the performance measure for the total error
- $m$ is the number of splits

The choice of the number of pieces in a split, i.e. the value of $k$, is delicate. We cannot choose $k$ too large, as the pieces become so small ($n_j$ becomes small) that we cannot distinguish the tops of the oscillations from the trend. On the other hand, if we choose $k$ too small a full "period" of the trend might fall in one piece, so that the large differences between the actual water level and fitted trend cancel out. We decide to keep the length of the piece, $n_j$ for $j \in \{1, ..., k-1\}$ constant (only the last piece, which contains the remaining values, differs in size). This implies that we vary the amount of pieces as we change the length of the fit range. After we experimented with various lengths of pieces we find that $n_j = 1500$ seconds is a reasonable length. This means that number of pieces in a split is given by

$$k = \lceil \frac{\text{fitrange}}{1500} \rceil \tag{6.24}$$

$perfdtrnd$ has to stay low. But how low? We should indicate detrends that are obviously not done adequately (there are clear patterns of a trend in the detrended plot) as wrong, in particular detrends of a peak that largely cross the inflection point of that peak in the first model (see Section 7.2). On the other hand, we should accept detrends within the inflection points. Amongst different peaks, there are quite some differences of the value $perfdtrnd$ takes for wrong detrends and acceptable detrends. As a first idea, we might define a condition for acceptable detrends relative to the peak. However, the created model, and as a part also the detrend should function for any future peak, not just a specific one. Therefore, we should define the condition on the performance measure independent of the peak. After some experimenting, we establish a bound for $perfdtrnd$. For detrend to be considered good, it should hold that

$$perfdtrnd < 1.3 \tag{6.25}$$

### 6.5.3 Detrending model 1

We will discuss the general approach taken for the first model. The approach of detrending for the other models follows this approach, only with different trends than a parabolic trend that need to be detrended. As we mentioned in Section 6.5.2 the same formula for the trend is used in detrending as stipulated in the model. In the first model, we model the trend as a second degree polynomial, meaning that we fit a second degree polynomial to detrend the data. As an example, we consider the peak on 30 August (this the peak with the highest top of the dataset) with a fitting range of 1 hour and the end point 5 minutes before the maximum. The plot of the fitted trend is shown in Figure 6.1.

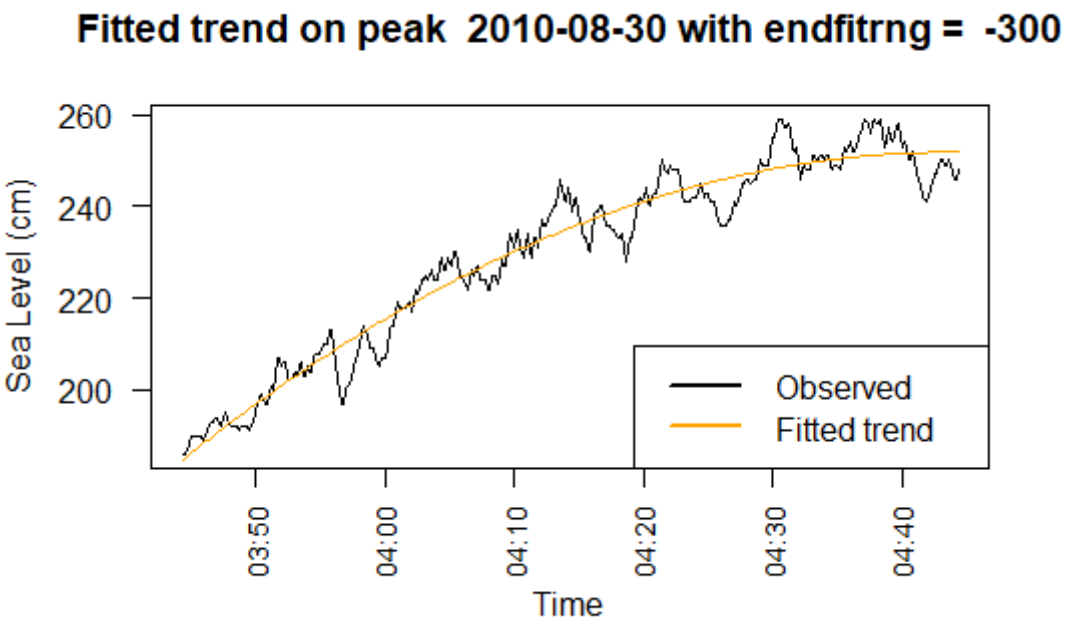


Figure 6.1: Example of a fitted trend with the peak of 30 August ($fitRng = 3600$, $endfitrng = -300$)

We obtain the detrended data by subtracting the fitted values of this model from the original measurements. If we apply this to the example peak of 30 August with a fit range of 1 hour, we obtain Figure 6.2.

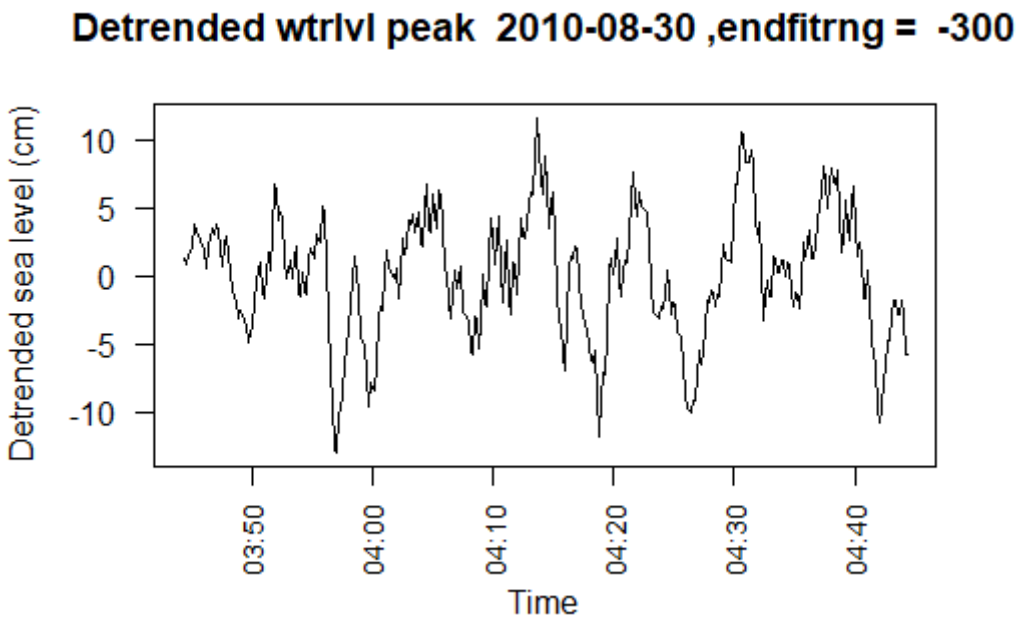


Figure 6.2: Example of the detrend with the peak of 30 August ($fitRng = 3600$, $endfitrng = -300$)

The corresponding value of the performance measure is

$$perfdtrnd = 0.5005507 \tag{6.26}$$

is lower than 1.3, indicating an acceptable detrending.

What does wrongly detrended water level look like? Again we consider the peak on 30 August with the same end point of the fit range, but now with a fit range of 2 hours. This way, the inflection point of the peak is amply crossed. The result of the fitted trend can be seen in Figure 6.3.

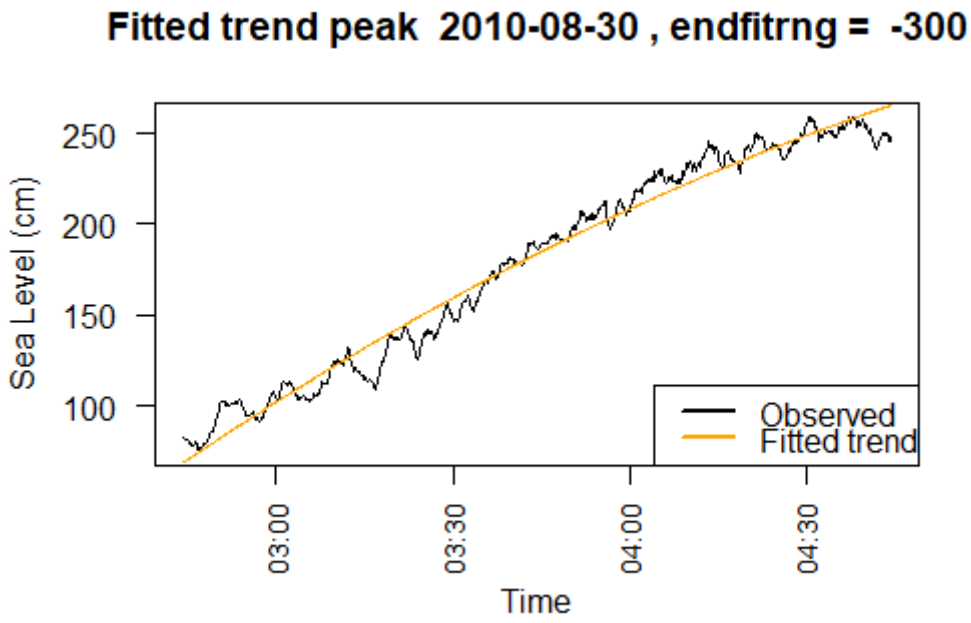


Figure 6.3: Example of a fitted trend with the peak of 30 August ($fitRng = 7200$, $endfitrng = -300$)

As we see, the second degree polynomial does not adequately capture the trend of the sea level. A more extensive inspection indicates that perhaps a third degree polynomial would have been a better fit for the trend.

We obtain the detrended data in the same way as before, and is plotted in Figure 6.4.

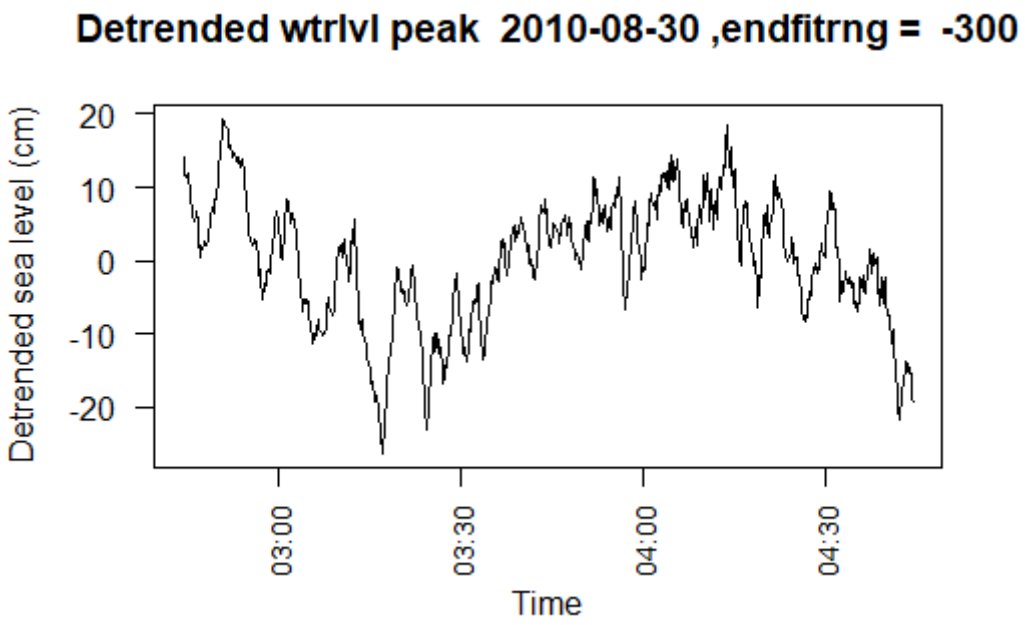


Figure 6.4: Example of the detrend with the peak of 30 August ($fitRng = 7200$, $endfitrng = -300$)

Graphically, we can clearly see that a trend is still present in data. This should also be indicated by the performance measure. Indeed, a value of

$$perfdtrnd = 9.11318$$

is larger than 1.3, indicating that the detrend is bad.

### 6.5.4 Possible issues

Because in this model we consider the water level to be a pure combination of the tide and three short-term oscillations, this implies that the trend is fully explained by the tide. Therefore we will use these words interchangeably within this model. We thus consider the trend to be a sinusoidal wave, and the second degree polynomial form in the model is essentially a Taylor approximation. In that light, it makes sense that the fitted parabolic trend that we subtract should be curved downwards and not upwards (that is to say, the extremum should be a maximum and not minimum). As it turns out, some fitting ranges will lead to upwards curved parabolic trends. One remedy would be that we do not allow the corresponding fitting ranges and somehow remove them within the predictions. However, this is very cumbersome work. Therefore, our approach will be to ignore the strange behaviour, and rely on the performance measures for the detrending and the prediction to filter out these fitting ranges. Moreover, if both performance measures would yield good results on some of these fitting ranges, then we predict that that fitting range is considered good in spite of possible unnatural detrending.

## 6.6 Speed of convergence fitting

One of the requirements of the prediction [R3], mentioned in the problem description (see Section 2), requires the predictions, and thus also the fitting, to be produced within one minute. The first model adheres to this requirement, the non-linear squares optimization is clearly solved within a minute. For the second model, the situation is different. In Section 6.4 we have seen that starting values for the height, the shift and the scale of the trend and the phases of the oscillations are supplied by a grid. The starting values of the height lie between 160 and 210, with time step of 1, giving 51 values. The starting values of the shift lie between 0 and 0.01, with time step 0.0005, giving 21 values. The starting values of the scale lie between 60000 and 220000, with time step of 20000, giving 9 values. Finally, the starting values of the phases lie between 0 and $\frac{8\pi}{5}$, with time step $\frac{2\pi}{5}$, giving 5 values. Thus, if all combinations of starting values are tried, $51 \cdot 21 \cdot 9 \cdot 5^3 = 1204875$ non-linear

squares optimizations need to be performed. It is no surprise that the time this takes largely exceeds the one minute-requirement.

Thus another approach is needed. Instead of optimizing all parameters at once, the optimization is splitted into two parts. First, a model with solely the trend, so without the short-term oscillations is fitted. That is, the non-linear squares optimization is performed on the model

$$Y_i = \beta_0 + \lambda^2 \left( \frac{t_i}{B} + \tilde{A} \right) e^{-\lambda\left(\frac{t_i}{B} + \tilde{A}\right)} + \varepsilon_i \tag{6.27}$$

Then, the estimates of the non-linear parameters of this model, i.e. the coefficients of the non-linear parameters in the regression object, are used as starting values for these same parameters, but now for the non-linear squares optimization of the full model (as described in Formula 5.14). In this optimization, a grid of starting values for the phases of the short-term oscillations are supplied. The frequencies of the short-term oscillations still have one starting value, the three most dominant frequencies found by Fourier analysis on the detrended water level. Using this splitted two-step optimization, the total time of the fitting takes approximately 4 minutes, which means that requirement [R3] is not yet fulfilled, but it is a large improvement over the original optimization, which took already more than 2 hours.

## 6.7 Static modelling model 1

In Chapter 4.4 we found two peaks that seemed to be well modelled by the second model than the first, parabolic model: the peak on August 29 and on March 1. Therefore, it seems natural to compare the two models on these two peaks. We also include one peak where the second model did not seem an obvious choice, to compare with the first two peaks. The extra peak we include is August 30. The approach is as follows: we plotted the fitted values and the observed values for the three peaks. We took a fit range of 20 minutes, with the top exactly located at the middle of the fit (i.e. $endfitRng = 600$). In this Section, we cover the first, parabolic model. In the Section 6.8, we cover the second model.

The plots of the static fits of the first model for the considered peaks can be seen in Figure 6.5, Figure 6.6 and Figure 6.7. Considering the plots, there seem to be a clear hierarchy on the quality of the fits: the peak of August 29 seem to fit the best, then the peak of August 30 and lastly the peak of 1 March. However, this supposed difference in performance can partly be explained by the different range of the water level on the two days. On 1 March the range was approximately 10 centimeter while on 29 August and 30 August the range was approximately 25 centimeters.

We can more easily compare the quality of the different fits if we quantify the quality in an error term. To calculate this error, we use the first layer individual error (discussed in Chapter 3, see Formula 3.4). Notice that here we apply the performance measure on the fitted values rather than on the predicted values. In Chapter 8 we will use the performance measure for its main purpose: we will apply it on the predicted values.

We see that the errors of two days are quite close to each other: $\varepsilon^{(j)}_{indv} = 1.19$ for the peak on 1 March against $\varepsilon^{(j)}_{indv} = 1.29$ on 29 August. The error on 30 August is approximately one centimeter higher: $\varepsilon^{(j)}_{indv} = 2.41$. It is interesting to see that two of the three errors are very close to 1 centimeter, so that at least statically the fits are close to demands of the Dutch Ministry of Infrastructure and Water Management of 1 centimeter error. Of course, we have to see whether this 1 centimeter error can be maintained by the dynamical fitting and the predictions.

Finally, we discuss the estimation of the top. On the peak of 29 August, the top is estimated quite good by the model. The time is almost exactly correct, and the location, i.e. the water level, differs approximately by 3 centimeter. On 1 March, there is not one unique top, but rather there are 2 times 2 tops close to each other to distinguish. Both the time of these pairs are identified correctly by the first model, the water level of the top of the model is 1 centimeter lower than the actual tops.

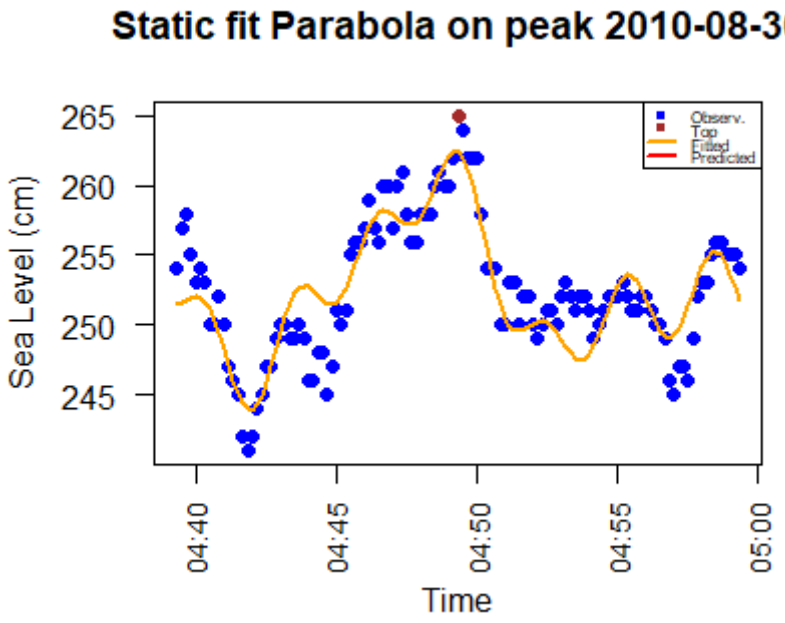


Figure 6.5: One static fit of model 1 10 minutes before and 10 minutes after top, peak of 30 August 2010

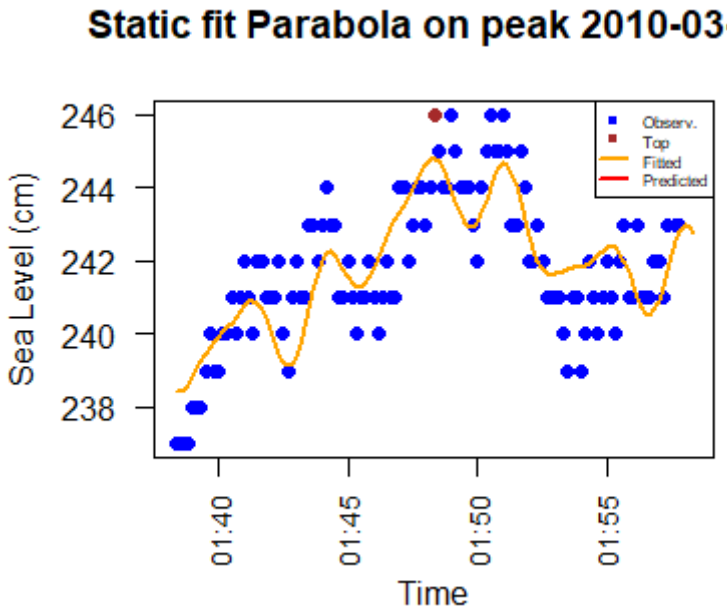


Figure 6.6: One static fit of model 1 10 minutes before and 10 minutes after top, peak of 1 March 2010

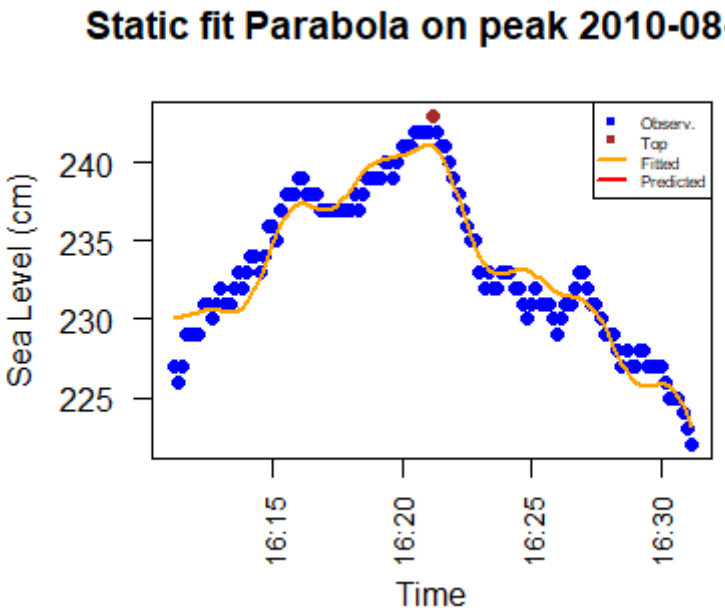


Figure 6.7: One static fit of model 1 10 minutes before and 10 minutes after top, peak of 29 August 2010

## 6.8 Static modelling model 2

With the second model we see a similar graphical effect as with the first model: again, considering the plots, it seems as if the peak of 1 March has a better fit than 29 August. If we again calculate the average over the absolute values of the errors of the fitted values, we do see an increased difference between 1 March and 29 August: $\varepsilon_{indv}^{(j)} = 1.27$ for the peak on 1 March and $\varepsilon_{indv}^{(j)} = 1.71$ on 29 August. Both errors again are quite close to demands of the Dutch Ministry of Infrastructure and Water Management of 1 centimeter. Again the error on 30 August is quite high: $\varepsilon_{indv}^{(j)} = 3.20$.

For the second model, we see similar behaviour as the first model for the tops. Again the location of the tops are identified by the model, although the fitted top water level lies lower than the actual water level.

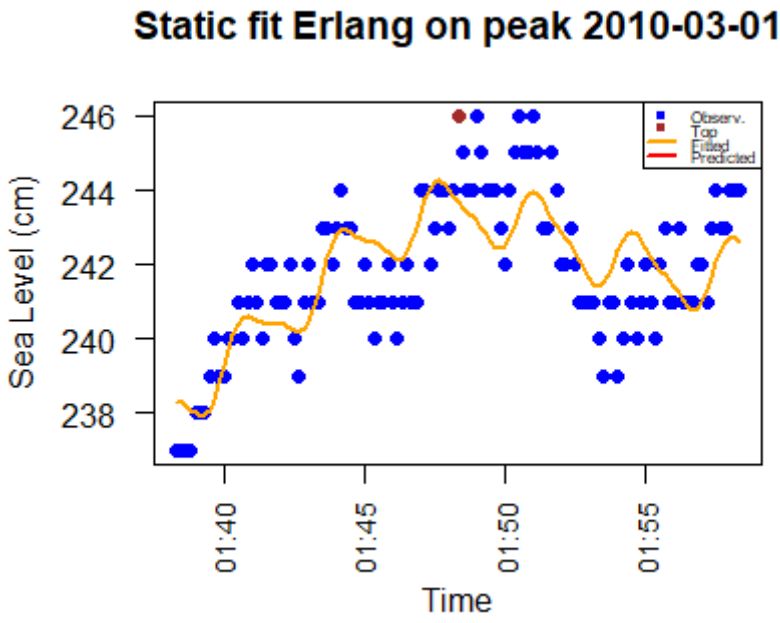


Figure 6.8: One static fit of model 2 over full peak, peak of 1 March 2010

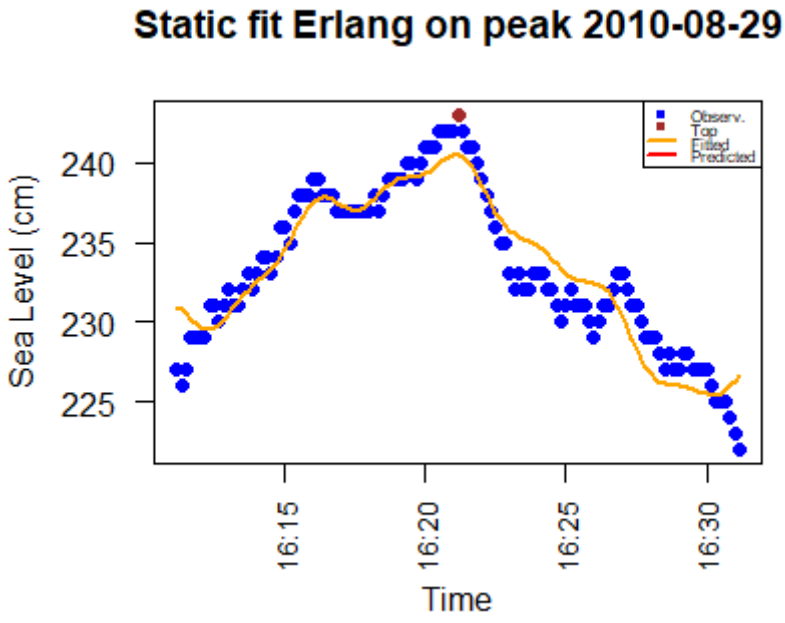


Figure 6.9: One static fit of model 2 over full peak, peak of 29 August 2010

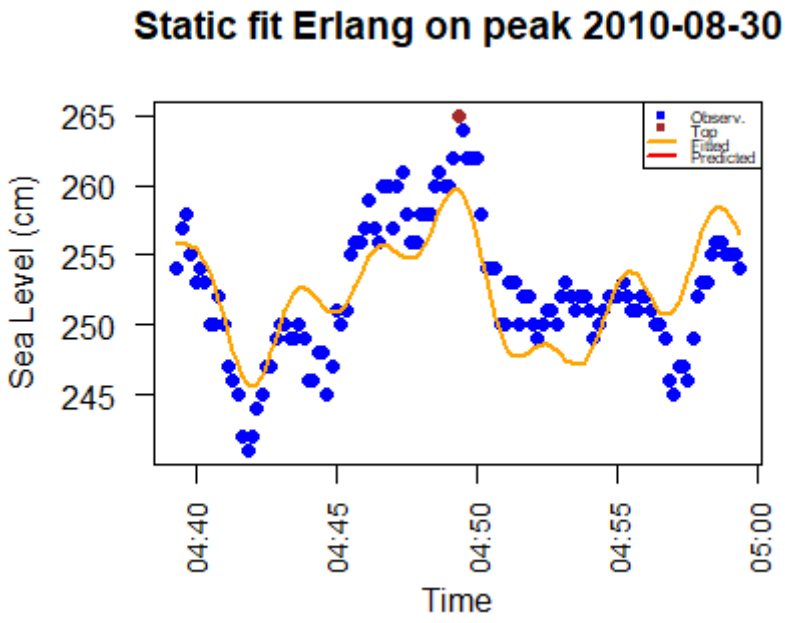


Figure 6.10: One static fit of model 2 over full peak, peak of 30 August 2010

## 6.9 Comparison first and second model

If we compare the two models, we first of all see that the first model leads to smaller errors for all peaks than the second model. If we take a closer look, we see that in terms of first layer error of the fit, the models are quite comparable considering the peak of 1 March and 29 August. The second model is thus able to statically model the peaks it was meant to model. It can do so almost equally well as the first model, but not better. We have to see if this situation changes for the dynamical fit and the predictions. We also see that the peak meant for verification contributes to a relatively high error for both models. With a difference of more than 1 centimeter we tend to say that the second model indeed is significantly worse than the first model, as we expected from the plots in Chapter 4.4. In terms of ability to find the locations of the top, both models perform almost equally good.

## 6.10 Conclusion

In this Chapter, we have statically fitted the parabolic and the Erlang model we build up in Chapter 5 to investigate the general "goodness-of-fit" of the model and test (the form of) the model itself. Two ways were discussed to reduce the amount of non-linear parameters. The first one relies on expressing the rate parameter $\lambda$ in terms of the independent variable, the shift parameter $\tilde{A}$ and the scale parameter $B$ by using the condition that the derivative should be equal to zero at the top:

$$\widehat{\lambda} = \frac{1}{\frac{x}{B} + \tilde{A}} \tag{6.28}$$

The number of non-linear parameters in the short-term oscillations can also be reduced: By rewriting

$$a \sin(\omega t_i + \rho) = \alpha \sin(\omega t_i) + \beta \cos(\omega t_i) \tag{6.29}$$

where

$$\alpha = a \cos(\rho) \text{ and } \beta = a \sin(\rho) \tag{6.30}$$

we transformed the original form with two non-linear parameters and one conditionally linear parameter to a form with two conditionally linear parameters and one non-linear parameter. In practice, only the first reduction is used. The second model then turns into

$$Y_i = \beta_0 + \left(\frac{1}{\frac{t_{max}}{B} + \tilde{A}}\right)^2 \left(\frac{t_i}{B} + \tilde{A}\right) e^{-\left(\frac{1}{\frac{t_{max}}{B} + \tilde{A}}\right)\left(\frac{t_i}{B} + \tilde{A}\right)} + a_1 \sin(\omega_1 t_i + \rho_1) + a_2 \sin(\omega_2 t_i + \rho_2) + a_3 \sin(\omega_3 t_i + \rho_3) + \varepsilon_i \tag{6.31}$$

The form of the first model stays as given in Formula 5.6. To supply the starting values for the models the height $\beta_0$, the shift $\tilde{A}$, the scale $B$ of the Erlang distribution and for the phases $\rho$ a seperate range of starting values is given. Estimates of the frequencies $\omega$ are found by a Fourier analysis on the detrended data. To check whether the data is detrended adequately, we developed a performance measure:

$$perfdtrnd = \max_{\ell \in \{1,\ldots,m\}} \max_{j \in \{1,\ldots,k\}} \left| \frac{1}{n_j} \sum_{i=1}^{n_j} (y_i^\ell - \widehat{y}_i^\ell) \right| \tag{6.32}$$

For a detrend to be considered "good", it should hold that

$$perfdtrnd < 1.3 \tag{6.33}$$

Because in the second model up to six parameters are optimized using a given range, the grid of starting values becomes extremely large, so that the algorithm is not able to terminate within one minute. Therefore, the trend is fitted first, and then these estimates are used in fitting the full model. The time is taken relative to the top of the peak, i.e. the "origin" is the top of the peak.

Finally, we discuss the static fits of the two models. The first thing to notice is that the first model leads to smaller errors for all peaks than the second model. We also see that in terms of first layer error of the fit, the models are quite comparable considering the peak of 1 March and 29 August, the peaks that created the second model. The second model is thus able to statically model the peaks it was meant to model. It can do so almost equally well as the first model, but not better. The "verificational" peak contributes to a relatively high error for both models. With a difference of more than 1 centimeter in error we tend to say that the second model indeed is significantly worse than the first model for this peak, as we expected from the plots in Chapter 4.4. In terms of ability to find the locations of the top, both models perform almost equally good. Both models identify the location of the tops almost correctly, and the height of the top a bit higher than predicted, on average 3 centimeters.

# Chapter 7

# Dynamic model fitting

In this chapter, our main aim is to dynamically fit the first and second model we statically fitted in Chapter 6. Largely, the dynamic fit follows the same approach as the static one. The differences are discussed in Section 7.1

Because the trend is modelled locally, the models have a certain region or time frame where they are expected to fit "reasonably". In Section 7.2 we identify and practically find the starting point of this region for the first model (the so-called "boundary points of the model").

Next to that, in static fitting we know the location and height of the top beforehand, since we considered the full peak at once. During dynamical fitting, we mimic the real-time situation, so that we do not know the top of the peak before it occurs. To get a sense of the location of the top before it occurs, we derive an estimator for the location of the top in Section 7.3.

One of the main differences with static fitting (which is also discussed in Section 7.1) is that in dynamic fitting, the fitting range is not fixed anymore, but we can choose a size for it. In Section 7.4 we do experiments with different sizes to find the optimal size of the fit range.

Finally, in Sections 7.5 and 7.6 we dynamically fit the first and second model.

## 7.1 Similarities and differences with static fitting

As mentioned, the dynamical fitting process largely follows the approach taken in the static fitting process. There are, however, some differences with static fitting that we need to discuss. For instance, because the static fit encapsulated one fit, that needed to include the full peak, the fit range was more or less fixed. In dynamic fitting, however, we have multiple so-called "updates" of the fit, so that we can choose the size of the fit range.

For the dynamical fits, we do not know the location and height of the top as we did with static fitting. We do not express $\lambda$ in the shape parameter, scale parameter and independent variable as we did in Section 6.3. The second model thus again follows the original form (see Formula (5.14)). Hence, in addition to the starting values for the parameters mentioned in Section 6.4, we also need to supply a starting value for $\lambda$. This is done by supplying a grid of starting values.

Furthermore, the detrending process is entirely analogues to the detrending at the static modelling described in Section 6.5.2. Also similar to the static modelling, we again split the fitting of the second model into two parts: fitting solely the trend and fitting the full model.

## 7.2 "Boundary points" of the model

### 7.2.1 Origin and expected location

As we mentioned, the first model uses a parabola to model the tide. If we assume that the real tide can be modelled by a sine wave, then we can see this parabola as a second order Taylor approximation of the full tide. As usual, this approximation is only "reasonably" valid for a certain neighbourhood around the top. More precisely, if the sea level indeed purely were a combination of the tide and three oscillations, then we expect the fitted model to "reliably" represent the actual water height within the inflection points of the trend around the top. Also the inflection points would be symmetric around the top, having the same distance to the top.

In practice things are more complicated. From the plots in Section 4.4, we can see that other sources cause the trend after the top to deviate from the regular tide, so that there is no second inflection point present, or it lies closer to the top than the first one. Moreover, due to differences in individual character, the precise location of these inflection points will differ from peak to peak. Therefore, the inflection points are dependent on the peak we consider.

We notice that not only the predicting range, but also the fit range must in principle lie entirely beyond the first inflection point for an adequate detrending and therefore adequate prediction (see Section 6.5.3) To find the most distant, allowed starting point of the fitting range, we need the location of the first inflection point.

From the plots of the first six highest peaks (see Section 4.4) we can see that even when solely the top is above 3 meters, it can happen that the second inflection point, i.e. the left boundary point of the model is reached, before water level drops below 2.75 meters. In that case, there will be a time frame where the model cannot be used while we still need predictions, namely from the second inflection to the point the water level drops below 2.75 meters. When a larger portion of the peak is above 3 meters, we expect the point where the water level drops below 2.75 is to further shift to the right, while inflection point stays the same, so that this gap becomes bigger. To find the end point of the prediction range, and investigate how big the gap is of 'cautious' water level, we need the location of the second inflection point.

The details on how we exactly find the location of the first inflection point we will discuss in Section 7.2.2.

### 7.2.2 Finding first "boundary point" model

In Section 7.2.1 we discussed the notion of boundary points of the first model, as well as their importance for the model. In this section, we will describe the method on how to pin down the first (left) boundary point. When searching for both inflection points the distinction between theory and practice becomes obvious. Theoretically, inflection points are a well-defined concept, and their location is clear: it is the point where the derivative of a function has an extremum, maximum or minimum. While this theoretical point of view on the location will also be the way we find the location, in practice it will be harder to pin down the exact location, and instead our best result will be an approximate location. Both left and right inflection point suffer from this approximation, although the in Section 7.2 described asymmetry requires two different methods to trace down the location, so that also the extent of approximation differs for both.

Arguably the most intuitive approach on how we find the left boundary point is as follows: For a given, fixed step size, starting from the current moment in time we start calculating derivatives backward in time and keep a moving maximum. Once the calculated derivative is smaller than the moving maximum, we have found the first approximate location for the inflection point. We then repeat the same procedure within the two subsequent points (differing by first chosen step size), but now with a smaller step size, that is a divisor of the first step size. We keep repeating this procedure until the suggested location of the inflection point worsens. We investigated the step size for which this happens beforehand with the validated data and programmed the exact amount of iterations explicitly.

The main problem with this procedure is that we do not know beforehand when the water level exceeds the 2.75 meters.

If a larger part of the peak lies above 3 meters, then the point where the water level exceeds 2.75 meters to shift to the left, while the inflection point stays the same. As a result, the time frame between the inflection point at the possible starting measuring point decreases. If this time frame becomes too small, the above described method might not work anymore, as the supposed location for the inflection point will get stuck in one of the oscillations. Another with-drawl is that because of this varying time frame, we are not sure what initial step size to pick.

To overcome this problem, we come to a different approach. This approach is independent of the point the water level exceeds the 2.75 meters and thus the possible starting point. The main idea is that we approach the inflection point from below. First, we determine the location of the global minimum of the valley occurring before the peak. To do this we choose a step size and a tolerance, and move backwards in time from the top, calculating derivatives until the absolute value of the derivative has become smaller than the prescribed tolerance and the water level is low enough (lower than 150, somewhat arbitrary choice). Then, we divide the time frame from the global minimum to the current moment in time in two equally long pieces, and for both, we determine the slope. We select the piece with the highest slope, and repeat the procedure of dividing into two pieces and selecting the larger derivative, until the pieces have become odd. Indeed, we suppose that the location of the final maximal derivative is the inflection point.

The procedure of determining the global minimum just before the peak has a couple of parameters we can tweak. For instance, we can alter the step size, as well as the tolerance. Intuitively and informally speaking, we should choose values for these parameters so that "for as much peaks as possible the supposed minimum lies as close to the actual global minimum as possible". More formally, we should choose the vector of values of parameters that minimizes the sum of all differences between the supposed and actual global minima. We could implement a systematic approach that computes, for a large sequence of vectors of parameter values, the sum of the differences between the actual and the estimated global minima amongst all peaks. In that manner, we can choose the vector of parameter values that lead to the smallest sum of differences, and thus the best estimation of the global minima. We then could determine the actual global minimum by another, much more reliable method. This systematic approach is postponed to future research or extensions, and instead we find "reasonably good" parameter values by manual experimenting.

Manual experimenting indicates that the choice of both parameters is delicate. Too big step sizes are undesirable, as the location of the global minimum gets to rough. On the other hand the step size cannot be too small, as otherwise the search will get stuck in a local extremum: one of the tops or valleys of a short term oscillation. Considering the tolerance, a too big size is unwanted as the search will stop too soon, widely before actual minimum. Too small tolerance, on the other hand, can cause the search to overshoot: the first valley is ignored and the search goes through the second peak only to stop in the global minimum of one valley earlier.

## 7.3 Finding an estimate for the location of the top

Next to the value of the top, the location of the top, i.e. the moment in time the top occurs is of importance. The location of the top can be formally defined as

$$loctop_{peakNr} := \arg \max_{i \in peak_{peakNr}} y_i \tag{7.1}$$

where

- $loctop_{peakNr}$ is the location of the top of peak $peakNr$
- $peak_{peakNr} := peak[peakNr, \text{“}Ind.Begin\text{”}] : peak[peakNr, \text{“}Ind.End\text{”}]$, i.e. $peak_{peakNr}$ is the interval containing all indices from the begin of the peak till the end of the peak
- $y_i$ is the $i^{th}$ observed water level

This is the actual location of the top, which we can be easily determine if all observed water levels during a peak are known. For real-time predictions, this is not the case: we then only know the water levels until the present moment in time. To get an idea of the location of the top beforehand, we need an estimator for the location of the top.

### 7.3.1 First intuitive estimator

Arguably the most intuitive choice is that we mimic the formula of the actual location of the top:

$$\widehat{loctop}_{peakNr}^{(endfitRng)} := \arg \max_{i \in peak_{peakNr}} \widehat{y}_i^{(endfitRng)} \tag{7.2}$$

Here,

- $\widehat{loctop}_{peakNr}^{(endfitRng)}$ is the estimator of the location of the top of peak $peakNr$ of the fitted model with the end of the fit range positioned at $endfitRng$
- $endfitRng$ is the parameter controlling the end of the fit range
- $peak_{peakNr} := peak[peakNr, \text{“}Ind.Begin\text{”}] : peak[peakNr, \text{“}Ind.End\text{”}]$, i.e. $peak_{peakNr}$ is the interval containing all indices from the begin of the peak till the end of the peak
- $\widehat{y}_i^{(endfitRng)}$ is the $i^{th}$ predicted water level of the fitted model with the end of the fit range positioned at $endfitRng$

To see whether the estimates are close to the real top, we consider the bias:

$$b(\widehat{loctop}_{peakNr}^{(endfitrng)}) = loctop_{peakNr} - \widehat{loctop}_{peakNr}^{(endfitrng)} \tag{7.3}$$

Notice the dependence on $endfitRng$ in the estimator (and bias) of the location and in the predicted water level. This dependence is due to the dynamicness of the model: within a peak, we re-fit the model every minute (before a new decision is made), leading to new predictions of the water level. Because of this dependence, $\widehat{loctop}_{peakNr}^{(endfitRng)}$ is re-evaluated every time we re-fit the model.

| $endfitRng$ | $b(loctop_1^{(endfitrng)})$ | $b(loctop_2^{(endfitrng)})$ | $b(loctop_3^{(endfitrng)})$ |
|---|---|---|---|
| -600 | -580 | -620 | 6460 |
| -540 | -520 | -550 | 6450 |
| -480 | -740 | -460 | 6460 |
| -420 | -730 | -400 | 6270 |
| -360 | -720 | -340 | 6260 |
| -300 | -730 | -540 | 6460 |
| -240 | -720 | -510 | 6680 |
| -180 | -710 | -510 | -240 |
| -120 | -170 | -520 | -250 |
| -60 | -40 | -100 | 6640 |
| 0 | 20 | 20 | 6650 |
| 60 | 80 | 80 | 6660 |
| 120 | 140 | 140 | -20 |
| 180 | 200 | 200 | 170 |

Table 7.1: Bias of the estimator $\widehat{loctop}_{peakNr}^{(endfitRng)}$ from 10 minutes before up to 3 minutes after the top, for the three highest peaks applied to the first model

| $endfitRng$ | $b(loctop_4^{(endfitrng)})$ | $b(loctop_5^{(endfitrng)})$ | $b(loctop_6^{(endfitrng)})$ |
|---|---|---|---|
| -600 | -870 | -720 | -580 |
| -540 | 7030 | -530 | -520 |
| -480 | 6820 | -490 | 16780 |
| -420 | 7030 | -400 | -690 |
| -360 | 7040 | -340 | 11950 |
| -300 | 7050 | -280 | -550 |
| -240 | 7240 | -220 | -530 |
| -180 | 7230 | -160 | -540 |
| -120 | 7240 | -100 | -540 |
| -60 | -510 | -40 | -540 |
| 0 | -130 | 20 | 20 |
| 60 | 7630 | 80 | 30 |
| 120 | 16380 | 0 | 12400 |
| 180 | 7630 | -10 | 12400 |

Table 7.2: Bias of the estimator $\widehat{loctop}_{peakNr}^{(endfitRng)}$ from 10 minutes before up to 3 minutes after the top, for the fourth, fifth and sixth highest peaks applied to the first model

Considering Table 7.1 and Table 7.2, we see quite different behaviour of the estimates amongst the peaks. First of all, there is a distinction to be made in terms of the size of the bias. The bias at the first, second, fifth and sixth highest peaks is all of order of magnitude 100, while the third and fourth highest peak are of order of magnitude 1000. The estimates of some peaks are more affected by the refits than others. Put differently, some estimates are more sensitive to refits than others. An extreme example is the third highest peak, that is, the peak of March 1. The estimates of this peak change a bit over the refits, but the overall behaviour is that of a constant estimate over time. We see the same invariant behaviour at the fourth highest peak (peak of October 24) and the six highest peak (peak of September 26). The first and second highest peaks (August 30 and September 25) show more variation, although the bias only decreases significantly when only one or two refits are left before the actual top. On the fifth highest peak, of August 29, the situation is rather differently. Here, the bias seem to monotonously decrease to 0 along with $endfitRng$.

### 7.3.2 "Moving average" estimator

To make the estimates of all peaks more robust against the sensitivity of the refits, our attempt is to use the averages of the estimates. We thus propose a new estimator for the location of the top, based on the moving averages of the estimates:

$$\widetilde{loctop}_{peakNr}^{(endfitRng)} := \frac{1}{k} \sum_{j \in \{allcurendfitRng\}} \widehat{loctop}_{peakNr}^{j} = \frac{1}{k} \sum_{j \in \{allcurendfitRng\}} \arg \max_{i \in peak_{peakNr}} \widehat{y}_i^{(j)} \tag{7.4}$$

Here:

- $\widetilde{loctop}_{peakNr}^{(endfitRng)}$ is the moving average estimator for the location of the top of peak $peakNr$
- $allendfitRng$ is the vector containing all $endfitRng$ up to and including the current fit
- $k$ is the number of different $endfitRng$ considered up to the current one, i.e. the length of the vector $allcurendfitRng$
- $j$ is a dummy variable used to indicate the different instances of $endfitRng$
- $i$ is a dummy variable used to indicate the different fitted values within one fit

Again, instead of the estimates itself, we consider the bias of the estimates:

$$b(\widetilde{loctop}_{peakNr}^{(endfitrng)}) = loctop_{peakNr} - \widetilde{loctop}_{peakNr}^{(endfitrng)} \tag{7.5}$$

| $endfitRng$ | $b(loctop_1^{(endfitrng)})$ | $b(loctop_2^{(endfitrng)})$ | $b(loctop_3^{(endfitrng)})$ |
|---|---|---|---|
| -600 | -580 | -620 | 6460 |
| -540 | -550 | -585 | 6455 |
| -480 | -613 | -543 | 6457 |
| -420 | -643 | -508 | 6410 |
| -360 | -658 | -474 | 6380 |
| -300 | -670 | -485 | 6393 |
| -240 | -677 | -489 | 6434 |
| -180 | -681 | -491 | 5600 |
| -120 | -624 | -494 | 4950 |
| -60 | -566 | -455 | 5119 |
| 0 | -513 | -412 | 5258 |
| 60 | -463 | -371 | 5375 |
| 120 | -417 | -332 | 4960 |
| 180 | -373 | -294 | 4618 |

Table 7.3: Bias of the moving average estimator $\widetilde{loctop}_{peakNr}^{(endfitRng)}$ from 10 minutes before up to 3 minutes after the top, for the three highest peaks applied to the first model

| $endfitRng$ | $b(loctop_4^{(endfitrng)})$ | $b(loctop_5^{(endfitrng)})$ | $b(loctop_6^{(endfitrng)})$ |
|---|---|---|---|
| -600 | -870 | -720 | -580 |
| -540 | 3080 | -625 | -550 |
| -480 | 4327 | -580 | 5227 |
| -420 | 5003 | -535 | 3748 |
| -360 | 5410 | -496 | 5388 |
| -300 | 5683 | -460 | 4398 |
| -240 | 5906 | -426 | 3694 |
| -180 | 6071 | -393 | 3165 |
| -120 | 6201 | -360 | 2753 |
| -60 | 5530 | -328 | 2424 |
| 0 | 5015 | -296 | 2205 |
| 60 | 5233 | -265 | 2024 |
| 120 | 6091 | -245 | 2822 |
| 180 | 6201 | -228 | 3506 |

Table 7.4: Bias of the moving average estimator $\widetilde{loctop}_{peakNr}^{(endfitRng)}$ from 10 minutes before up to 3 minutes after the top, for the fourth, fifth and sixth highest peaks applied to the first model

Considering Table 7.3 and Table 7.4 we see that this estimator indeed is more robust to changes in the individual fits of the model: the bias deviates less as the top is approached (i.e. as $endfitRng$ approaches zero). Nevertheless, the bias is in all cases still larger than 5 minutes. For the third, fourth and sixth highest peak it is even larger than 40 minutes. With such a large bias, the estimator is not practically applicable. If we compare the results of this moving average estimator with the results of the original, first proposed estimator, we can explain this behaviour. For the largest group of peaks, namely all peaks except the sixth highest peak, the order of magnitude of the bias is the same for both estimators. The third and fourth highest peak, which are part of that group, thus seem to be peaks that have difficult estimable tops. As we can see in Table 7.2, for the original estimator, the sixth highest peak has two outliers of the bias: the estimate and the actual value differ by more than 3 hours. These outliers affect the moving average estimator of the sixth highest peak largely: the bias is increased largely at fits corresponding to the outliers, and decreases only slightly as the top is approached.

### 7.3.3 "Moving minimum" estimator

To overcome this unwanted effect, we somehow need to filter out these "bad" estimates. The quantity on which the filter applies needs to be independent of the actual location of the top as we do not know this beforehand. One such a quantity is the quality of the fit, more specifically the residual sum of squares. We expect that the quality of the separate estimates (again from Formula 7.2) depend on the residual sum of squares of a fit. More precisely, that the quality of the separate estimates is some increasing function of the residual sum of squares. In particular we expect that "bad" separate estimates come from "bad" fits, informally speaking. If we consider the estimates (coming from the first estimator) up to the current fit, and take the estimate corresponding with minimal residual sum of squares, it is very likely that we can avoid these "bad" estimates. In fact, we only need one "good" estimate with a lower residual sum of squares between furthermore "bad" estimates to avoid these "bad" estimates. This new proposed estimator thus can be thought of as a "moving minimum" estimator. In a formula, this estimator can defined as:

$$\widetilde{loctop}_{peakNr}^{(endfitRng)} := \{\widehat{loctop}_{peakNr}^{(endfitRng)} | \min_{endfitRng \in \{allendfitRng\}} (RSS(y_i^{(endfitRng)}))\} \tag{7.6}$$

$$= \{\widehat{loctop}_{peakNr}^{(endfitRng)} | \min_{endfitRng \in \{allendfitRng\}} (\sum_{i=1}^{n} (y_i^{endfitRng} - \widehat{y}_i{}^{(endfitRng)})^2)\} \tag{7.7}$$

Here:

- $\widetilde{loctop}_{peakNr}^{(endfitRng)}$ is the "moving minimum estimator" for the location of the top of peak $peakNr$
- $\widehat{loctop}_{peakNr}^{(endfitRng)}$ is the original, first proposed estimator for the location of the top of peak $peakNr$, defined in Formula 7.2
- $allendfitRng$ is the vector containing all $endfitRng$ up to and including the current fit
- $RSS$ is the residual sum of squares of a fit

| $endfitRng$ | $b(loctop_1^{(endfitrng)})$ | $b(loctop_2^{(endfitrng)})$ | $b(loctop_3^{(endfitrng)})$ |
|---|---|---|---|
| -600 | -580 | -620 | 6460 |
| -540 | -580 | -620 | 6450 |
| -480 | -580 | -620 | 6460 |
| -420 | -730 | -620 | 6270 |
| -360 | -720 | -620 | 6260 |
| -300 | -720 | -540 | 6260 |
| -240 | -720 | -510 | 6260 |
| -180 | -720 | -510 | -240 |
| -120 | -720 | -510 | -250 |
| -60 | -720 | -510 | -250 |
| 0 | -720 | -510 | -250 |
| 60 | -720 | -510 | -250 |
| 120 | 140 | -510 | -250 |
| 180 | 200 | -510 | -250 |

Table 7.5: Bias of the "moving minimum" estimator $\widetilde{loctop}_{peakNr}^{(endfitRng)}$ from 10 minutes before up to 3 minutes after the top, for the three highest peaks applied to the first model

| $endfitRng$ | $b(loctop_4^{(endfitrng)})$ | $b(loctop_5^{(endfitrng)})$ | $b(loctop_6^{(endfitrng)})$ |
|---|---|---|---|
| -600 | -870 | -720 | -580 |
| -540 | -870 | -530 | -520 |
| -480 | -870 | -490 | -520 |
| -420 | -870 | -490 | -520 |
| -360 | -870 | -490 | -520 |
| -300 | -870 | -490 | -520 |
| -240 | -870 | -490 | -520 |
| -180 | -870 | -160 | -520 |
| -120 | -870 | -160 | -520 |
| -60 | -870 | -160 | -520 |
| 0 | -870 | -160 | -520 |
| 60 | -870 | -160 | -520 |
| 120 | 16380 | -160 | -520 |
| 180 | 16380 | -160 | -520 |

Table 7.6: Bias of the "moving minimum" estimator $\widetilde{loctop}_{peakNr}^{(endfitRng)}$ from 10 minutes before up to 3 minutes after the top, for the fourth, fifth and sixth highest peaks applied to the first model

As we can see in Table 7.5 and Table 7.6, the "moving minimum" estimator seems to perform better than the "moving average" estimator on the third, fourth and sixth highest peaks (these are the peaks on March 1, October 24 and September 26). Indeed, these are the peaks with a relative high bias: more than 40 minutes. On the other hand, the "moving minimum" estimator performs slightly worse on the other peaks, where there are no occurrences of estimates with a relatively high bias. Finally, at the fifth highest peak there is an occurrence of an estimate with a relatively high bias (more than 4 hours), but with the lowest residual sum of squares, so that all subsequent fits take on that estimate. Luckily, this "bad" estimate occurs late in the updates, already after the actual top has occurred. Therefore, we consider this occurrence as not relevant.

### 7.3.4 Comparing and choosing between estimators

Although the first proposed estimator (Formula 7.2) is probably the most intuitive estimator, it lacks some desirable properties. Because it directly uses the maximum of the fit as estimate, it is very sensitive to the goodness of fit of that particular update of the fit. As a result, the estimates can deviate a lot between the updates, which makes it hard to pin down one estimate of the top. Also, "bad" estimates, that is estimates with a large bias and with a high residual sum of squares, cannot be filtered away.

As an alternative, we proposed the "moving average" estimator. The advantage of this estimator is that all previous "good" estimates are included in the calculation of the next estimate of the location of the top. The

downside, however, is that the same holds true for the "bad" estimates, and depending on the size of the bias, these "bad" estimates can have a large weight so that the estimate continues to have a large bias for a long time.

Another option we proposed was the "moving minimum" estimator. The advantage of this estimator is that "bad" estimates with a high residual sum of squares (which are expected to be the greatest portion of "bad" estimates) are filtered out from the estimate. A disadvantage is that also "bad" estimates with low residual sum of squares occur, and that at those occurrences, the "moving minimum" estimator exchanges a quite "good" estimate for a quite "bad" one. Another disadvantage is that also "good" estimates with high residual sum of squares occur, and that these are ignored completely.

In the end, we choose for the "moving minimum" estimator, as it is able to filter out the estimates with relatively high bias (more than 40 minutes).

## 7.4 Experimenting with the fit range

In this section, we will dynamically fit both models with different sizes of the fit range, to find the optimal size. We will compare four sizes: 900 seconds, 1200 seconds, 1500 seconds and 1800. Comparison of the different sizes is done by means of the layer one error $\varepsilon_{indv}^{(j)}$ described in Chapter 3. We present the results of each model in a table, and include one plot of the different fit ranges:

| *endfitRng* | **fitRng** = 900 | **fitRng** = 1200 | **fitRng** = 1500 | **fitRng** = 1800 |
|---|---|---|---|---|
| -1200 | 0.579 | 0.707 | 1.247 | 1.146 |
| -900 | 0.835 | 0.786 | 1.068 | 1.052 |
| -600 | 0.837 | 0.984 | 1.271 | 1.061 |
| -300 | 0.735 | 0.845 | 1.568 | 1.333 |
| 0 | 0.663 | 0.776 | 1.188 | 1.038 |
| 300 | 0.753 | 1.255 | 2.133 | 1.880 |
| 600 | 0.825 | 1.301 | 1.482 | 1.396 |

Table 7.7: Dynamic fits of four different sizes of $fitRng$ of **first** model over different parts of the peak with the corresponding error $\varepsilon_{indv}^{(j)}$

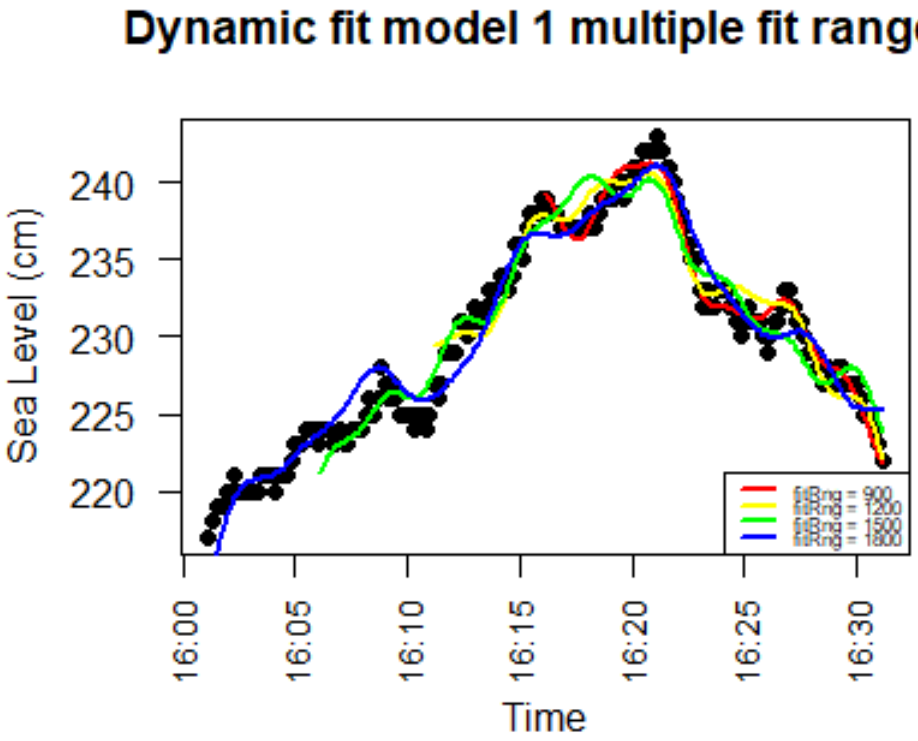


Figure 7.1: One dynamic fit of the **first** model with the four different sizes of $fitRng$. Here, $endfitRng = 600$

| $endfitRng$ | **fitRng** = 900 | **fitRng** = 1200 | **fitRng** = 1500 | **fitRng** = 1800 |
|---|---|---|---|---|
| -1200 | 0.594 | 0.744 | 1.154 | 1.162 |
| -900 | 0.915 | 0.843 | 1.205 | 1.048 |
| -600 | 0.906 | 0.957 | 1.342 | 1.219 |
| -300 | 0.779 | 0.926 | 1.543 | 1.286 |
| 0 | 0.635 | 0.806 | 1.211 | 1.167 |
| 300 | 0.878 | 1.491 | 2.283 | 2.020 |
| 600 | 0.750 | 1.173 | 1.731 | 1.714 |

Table 7.8: Dynamic fits of four different sizes of $fitRng$ of **second** model over different parts of the peak with the corresponding error $\varepsilon_{indv}^{(j)}$

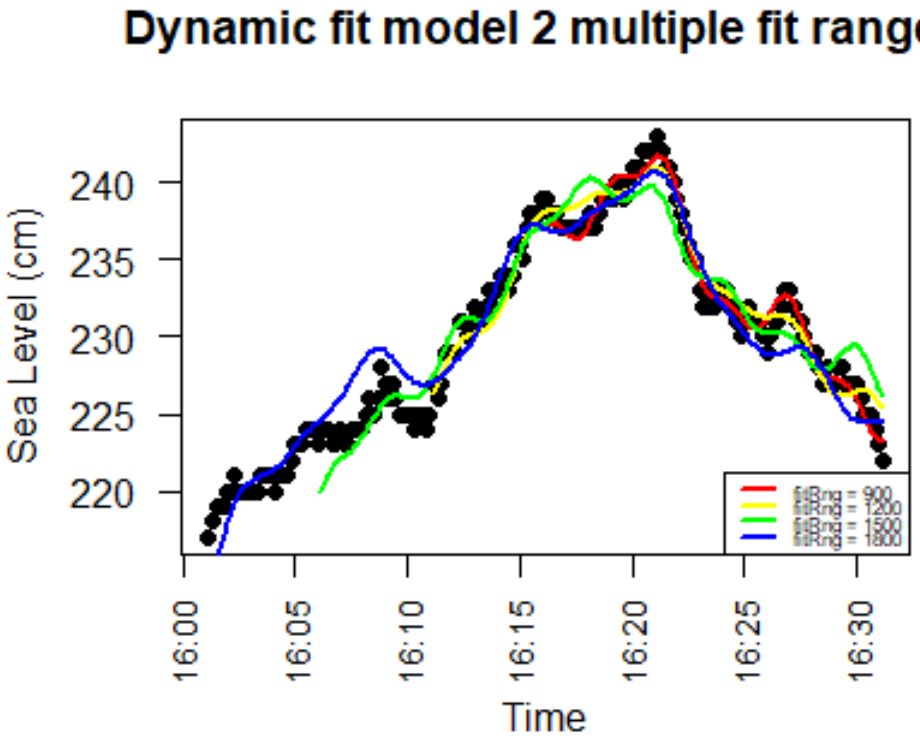


Figure 7.2: One dynamic fit of the **second** model with the four different sizes of $fitRng$. Here, $endfitRng = 600$

From the tables and the plots it can be seen that generally, a smaller size of $fitRng$ leads to better fits. Indeed, the smallest size $fitRng = 900$ leads to the best fits. However, too small fit ranges may also not be desirable, as the short-term oscillations are then not captured adequately. To ensure that these oscillations are well captured, we require that at least two periods of the short-term oscillations should be present in the fit range. Relying on the estimates of the Dutch Ministry of Infrastructure and Water Management, this means that the fit range should be at least $2 \cdot 545 = 1090$ seconds in size. Choosing the smallest next size we experimented with, we choose to use $fitRng = 1200$ seconds.

## 7.5 Dynamic fit model 1

### 7.5.1 Peak March 1

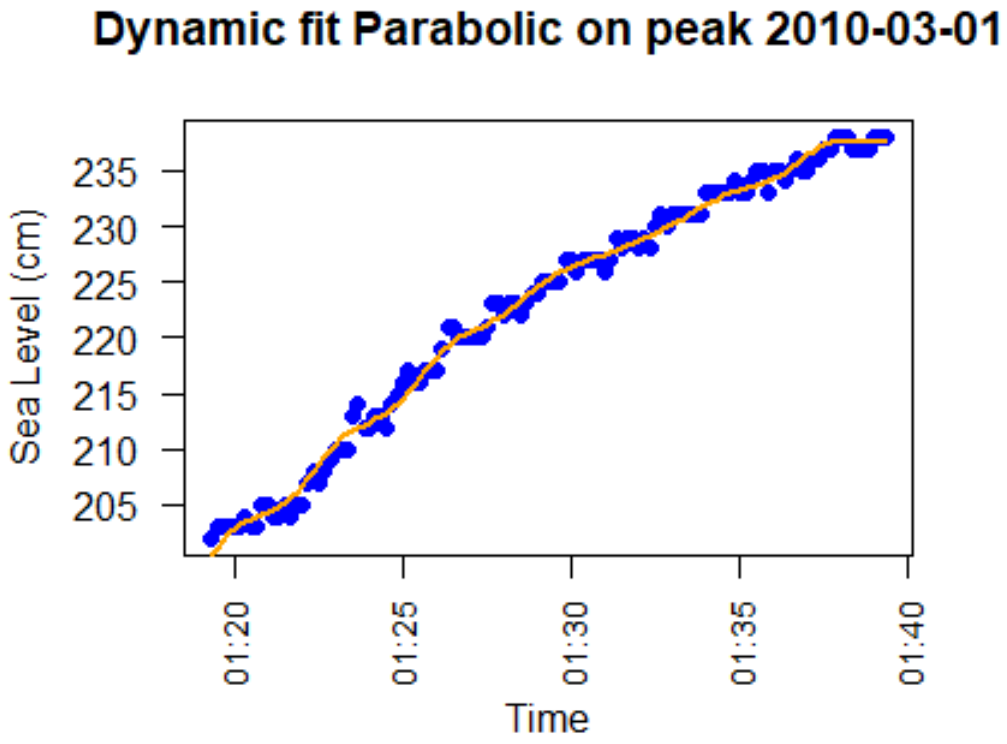


Figure 7.3: Dynamic fit first model peak March 1 10 minutes for top

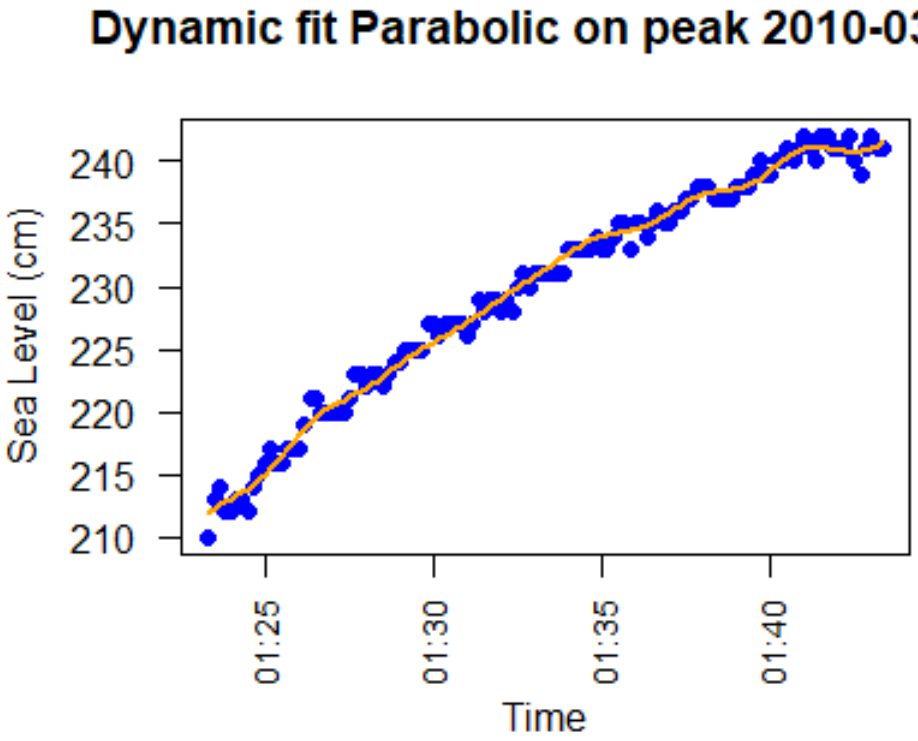


Figure 7.4: Dynamic fit first model peak March 1 5 minutes for top

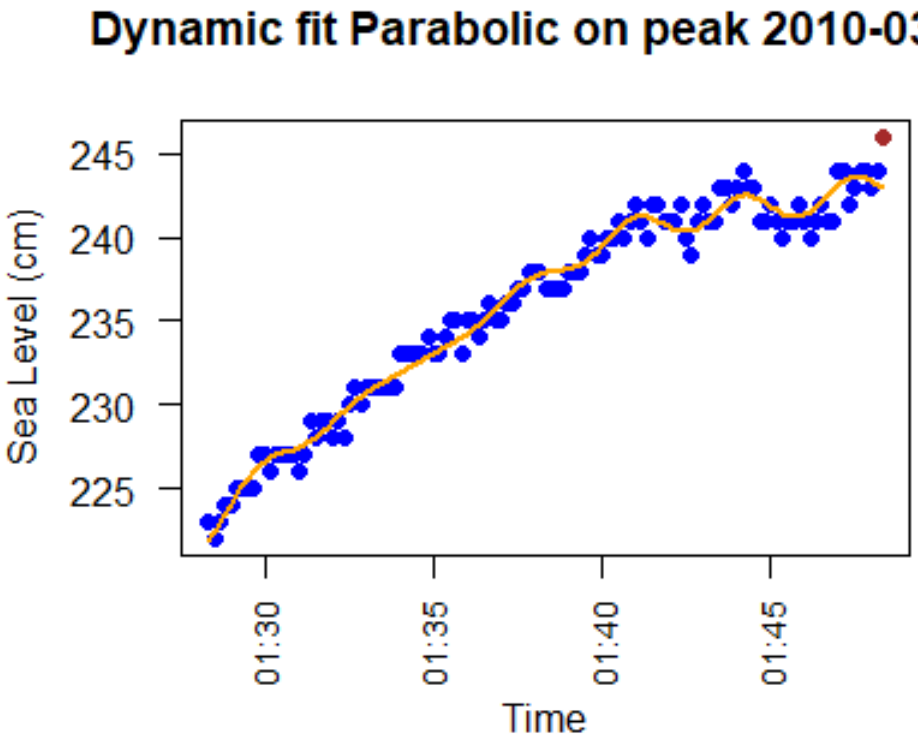


Figure 7.5: Dynamic fit first model peak March 1 0 minutes for top

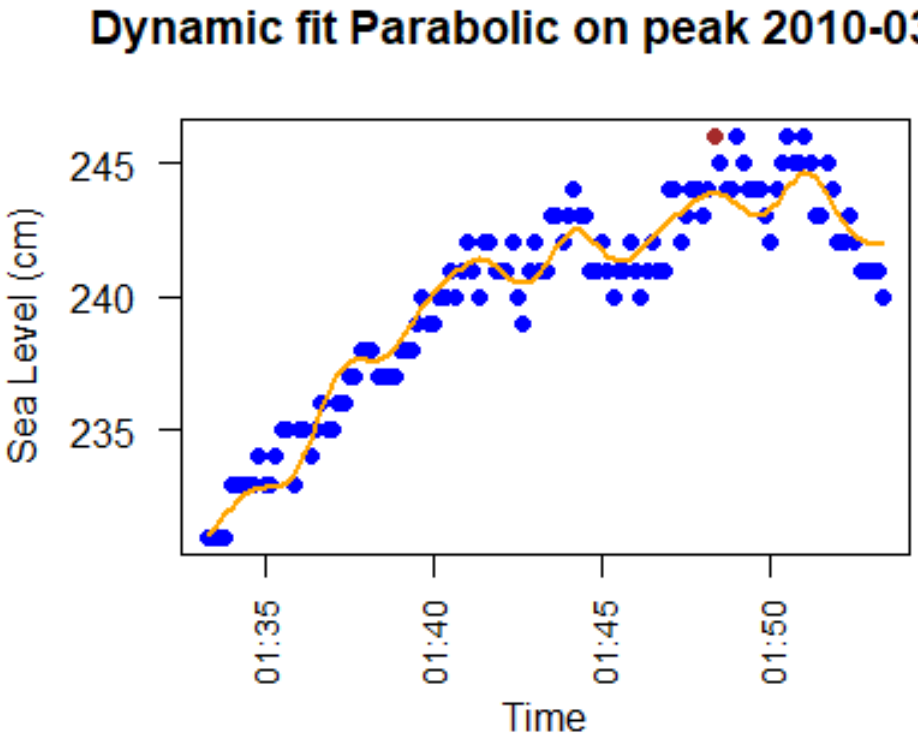


Figure 7.6: Dynamic fit first model peak March 1 5 minutes after top

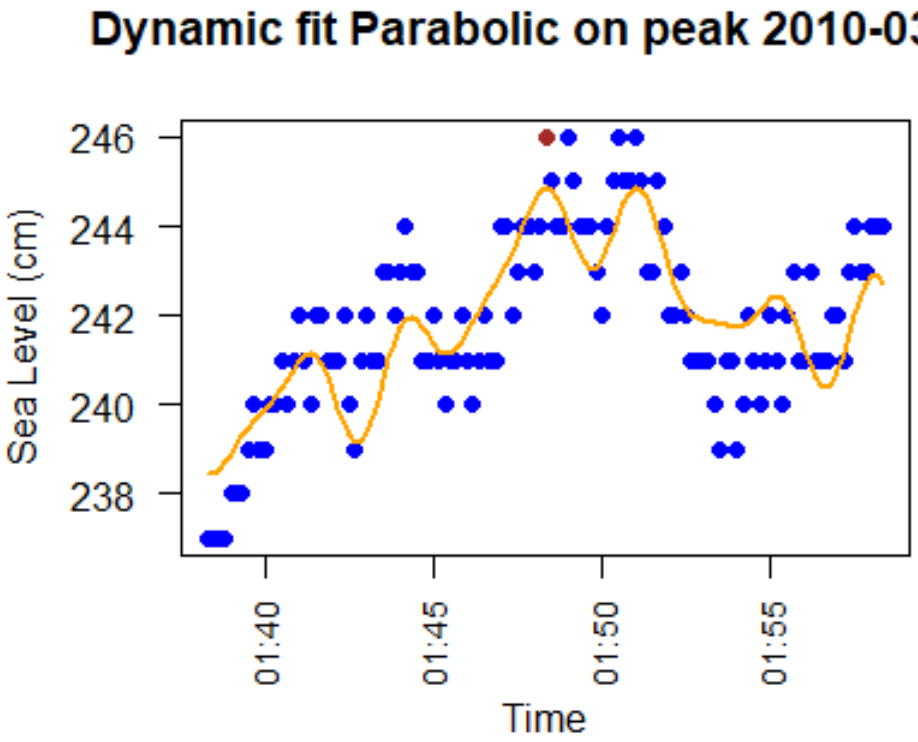


Figure 7.7: Dynamic fit first model peak March 1 10 minutes after top

| j | $endfitRng$ | $\varepsilon_{indv}^{(j)}$ |
|---|---|---|
| 1 | -600 | 0.859 |
| 2 | -540 | 0.839 |
| 3 | -480 | 0.815 |
| 4 | -420 | 0.756 |
| 5 | -360 | 0.746 |
| 6 | -300 | 0.799 |
| 7 | -240 | 0.790 |
| 8 | -180 | 0.735 |
| 9 | -120 | 0.732 |
| 10 | -60 | 0.809 |
| 11 | 0 | 0.820 |
| 12 | 60 | 0.886 |
| 13 | 120 | 0.903 |
| 14 | 180 | 0.897 |
| 15 | 240 | 0.925 |
| 16 | 300 | 0.969 |
| 17 | 360 | 1.300 |
| 18 | 420 | 1.008 |
| 19 | 480 | 1.060 |
| 20 | 540 | 1.055 |
| 21 | 600 | 1.177 |

Table 7.9: Dynamic fit of first model on peak March 1: individual error of 10 up to 10 minutes relative to the top

Combining the first layer individual errors $\varepsilon_{indv}^{(j)}$, we obtain the second layer error:

$$\varepsilon_{allavr} = 0.886 \tag{7.8}$$

### 7.5.2 Peak August 29

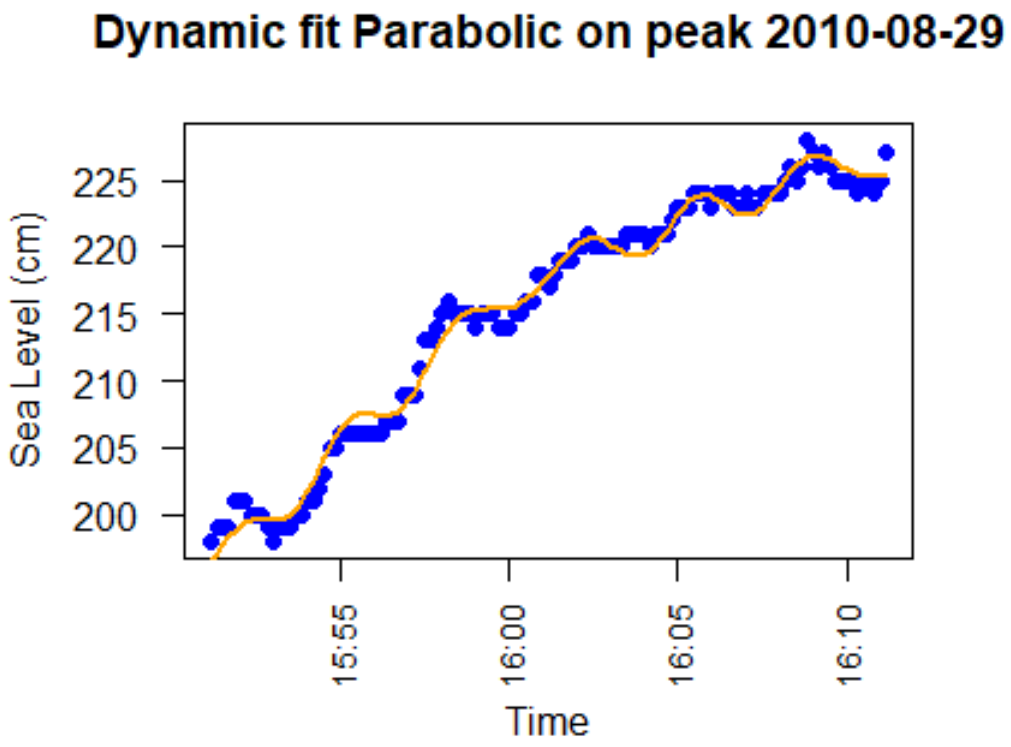


Figure 7.8: Dynamic fit first model peak August 29 10 minutes for top

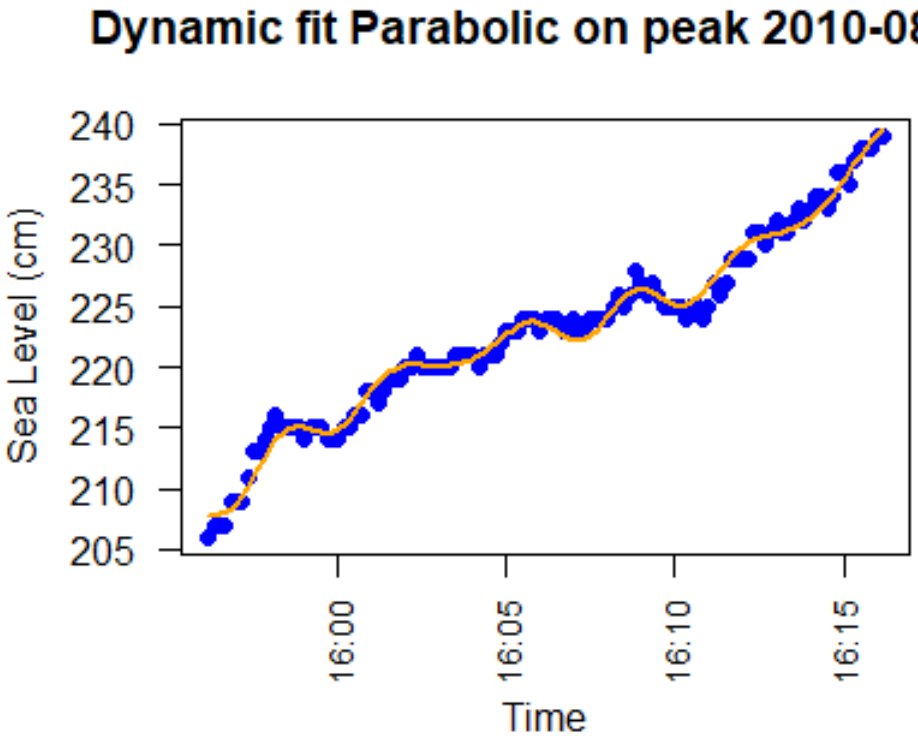


Figure 7.9: Dynamic fit first model peak August 29 5 minutes for top

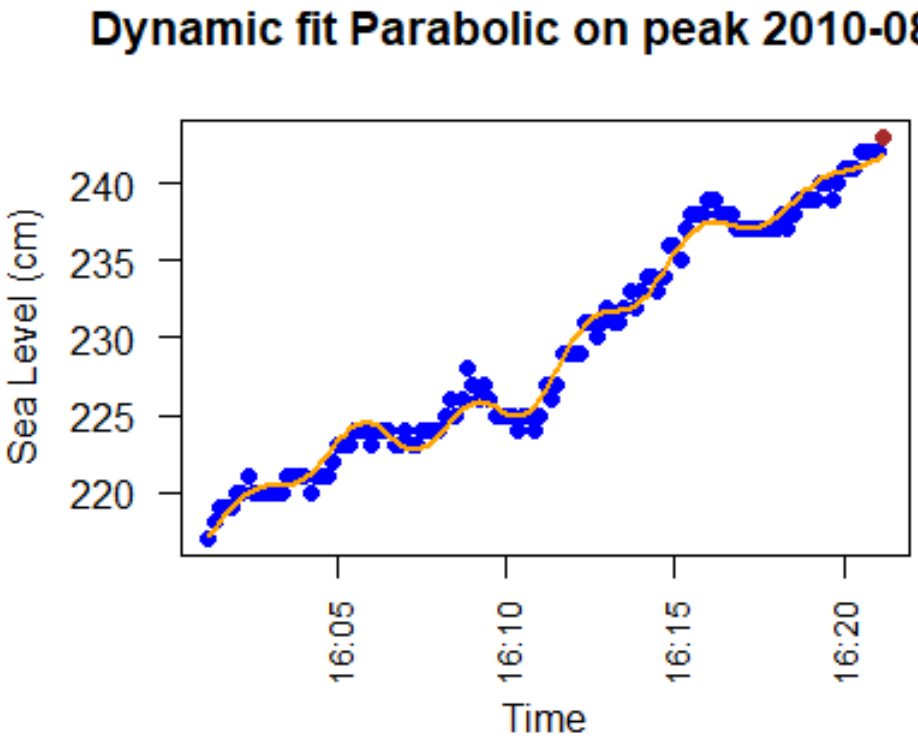


Figure 7.10: Dynamic fit first model peak August 29 0 minutes for top

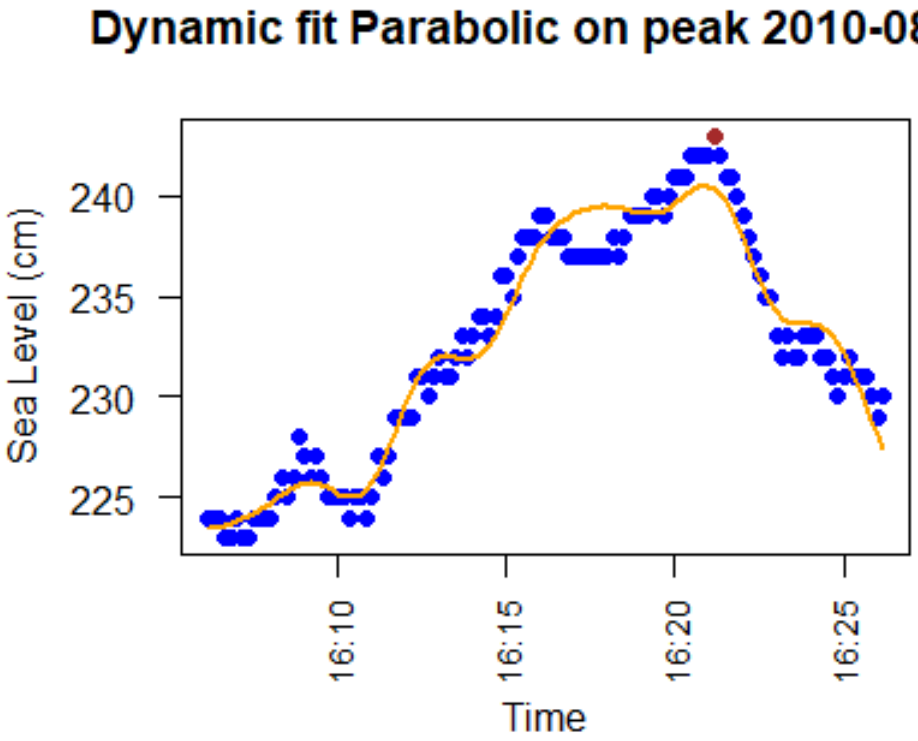


Figure 7.11: Dynamic fit first model peak August 29 5 minutes after top

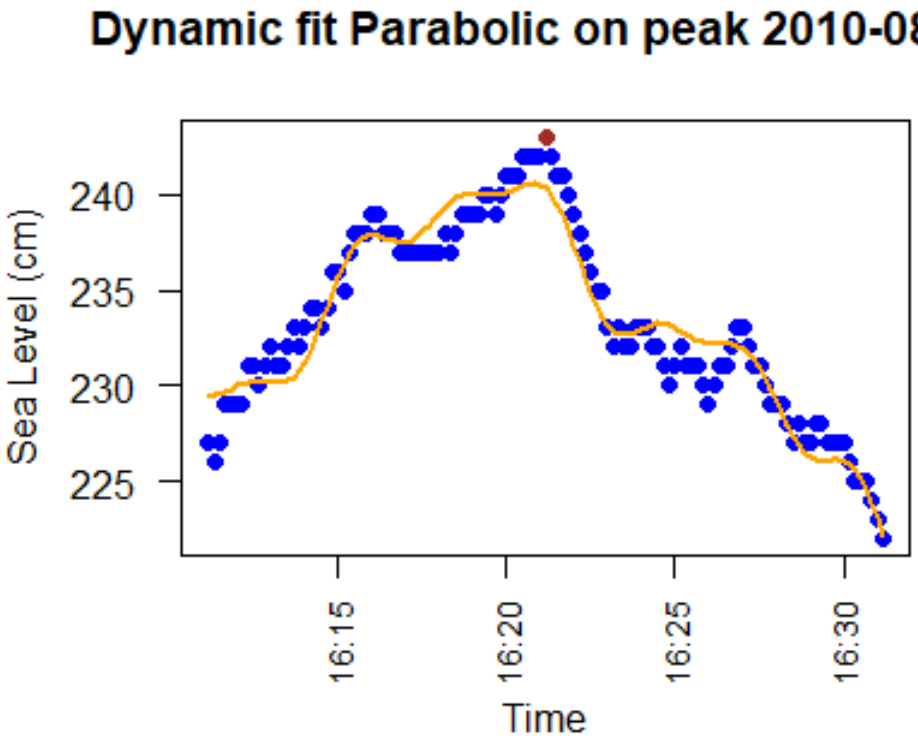


Figure 7.12: Dynamic fit first model peak August 29 10 minutes after top

| j | $endfitRng$ | $\varepsilon_{indv}^{(j)}$ |
|---|---|---|
| 1 | -600 | 0.984 |
| 2 | -540 | 0.887 |
| 3 | -480 | 0.829 |
| 4 | -420 | 0.898 |
| 5 | -360 | 0.896 |
| 6 | -300 | 0.845 |
| 7 | -240 | 0.963 |
| 8 | -180 | 0.781 |
| 9 | -120 | 0.789 |
| 10 | -60 | 0.778 |
| 11 | 0 | 0.776 |
| 12 | 60 | 1.029 |
| 13 | 120 | 1.192 |
| 14 | 180 | 1.252 |
| 15 | 240 | 1.158 |
| 16 | 300 | 1.255 |
| 17 | 360 | 1.717 |
| 18 | 420 | 1.750 |
| 19 | 480 | 1.497 |
| 20 | 540 | 1.619 |
| 21 | 600 | 1.301 |

Table 7.10: Dynamic fit of first model on peak August 29: individual error of 10 up to 10 minutes relative to the top

Combining the first layer individual errors $\varepsilon_{indv}^{(j)}$, we obtain the second layer error:

$$\varepsilon_{allavr} = 1.105 \tag{7.9}$$

### 7.5.3 Peak August 30

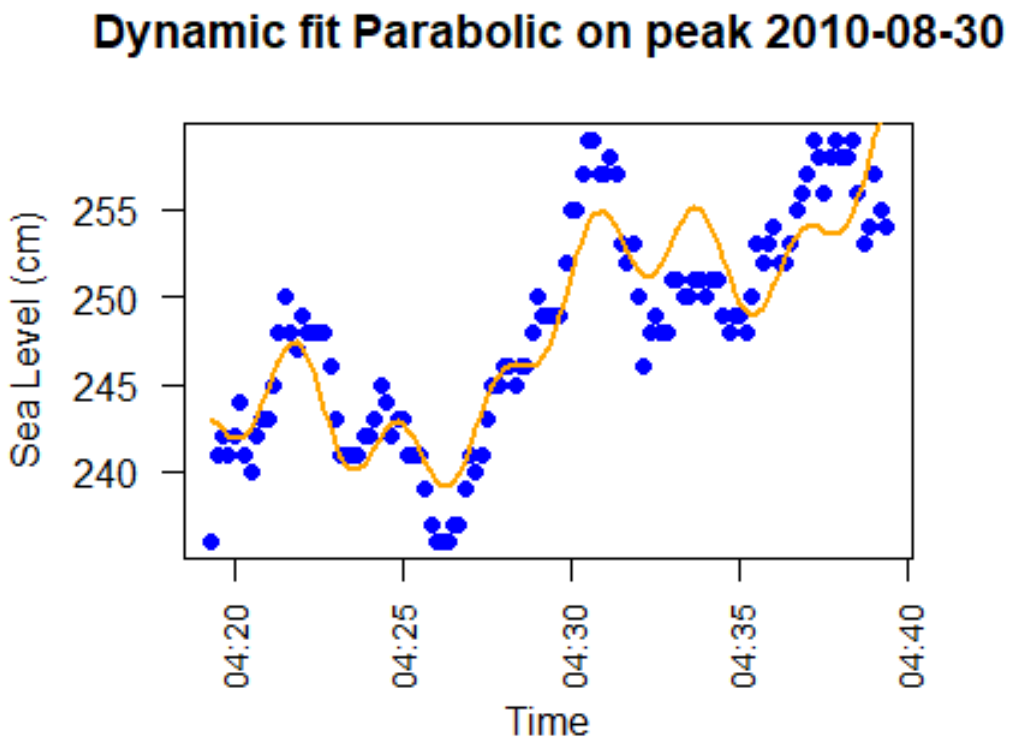


Figure 7.13: Dynamic fit first model peak August 30 10 minutes for top

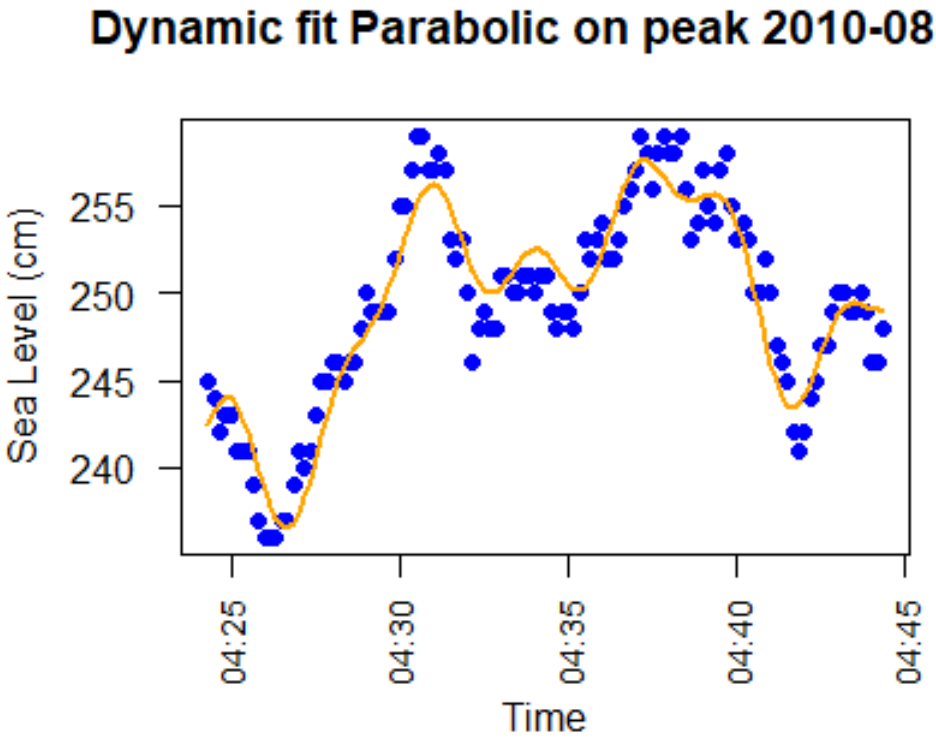


Figure 7.14: Dynamic fit first model peak August 30 5 minutes for top

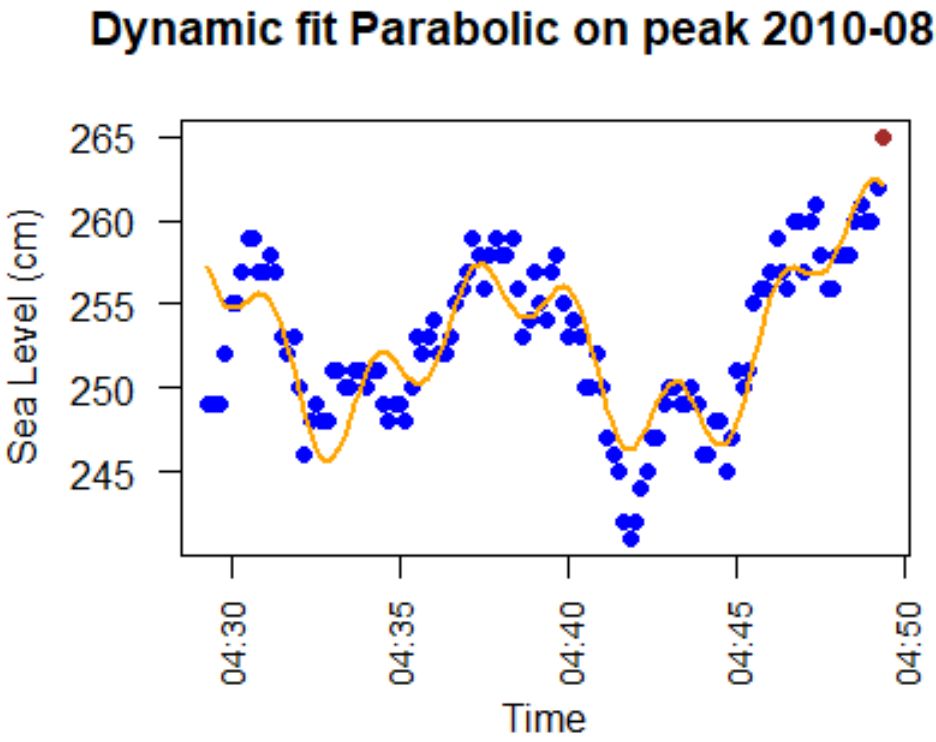


Figure 7.15: Dynamic fit first model peak August 30 0 minutes for top

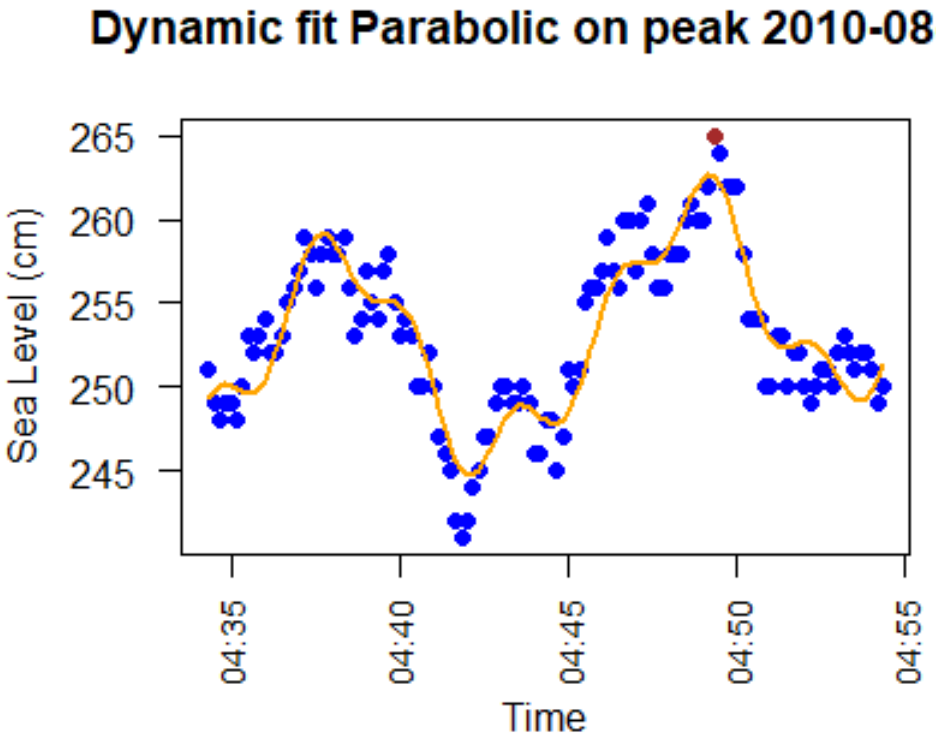


Figure 7.16: Dynamic fit first model peak August 30 5 minutes after top

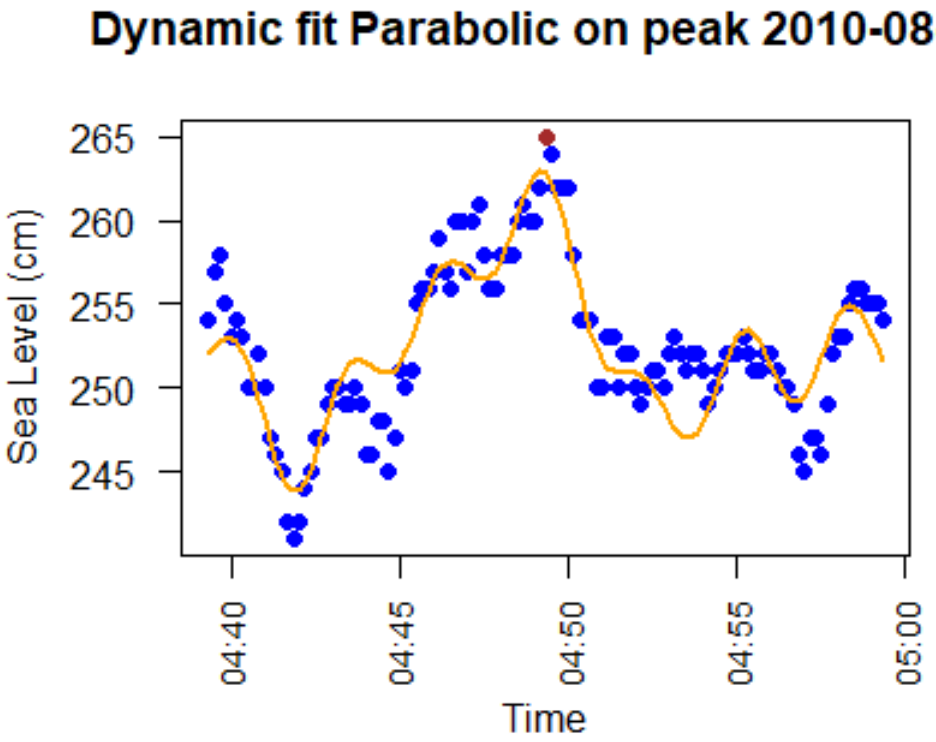


Figure 7.17: Dynamic fit first model peak August 30 10 minutes after top

| j | $endfitRng$ | $\varepsilon_{indv}^{(j)}$ |
|---|---|---|
| 1 | -600 | 2.770 |
| 2 | -540 | 3.084 |
| 3 | -480 | 2.776 |
| 4 | -420 | 2.674 |
| 5 | -360 | 1.871 |
| 6 | -300 | 1.899 |
| 7 | -240 | 2.278 |
| 8 | -180 | 2.723 |
| 9 | -120 | 3.317 |
| 10 | -60 | 3.229 |
| 11 | 0 | 2.435 |
| 12 | 60 | 1.978 |
| 13 | 120 | 2.124 |
| 14 | 180 | 2.169 |
| 15 | 240 | 2.052 |
| 16 | 300 | 1.935 |
| 17 | 360 | 1.965 |
| 18 | 420 | 1.876 |
| 19 | 480 | 2.580 |
| 20 | 540 | 2.661 |
| 21 | 600 | 2.328 |

Table 7.11: Dynamic fit of first model on peak August 30: individual error of 10 up to 10 minutes relative to the top

Combining the first layer individual errors $\varepsilon_{indv}^{(j)}$, we obtain the second layer error:

$$\varepsilon_{allavr} = 2.415 \tag{7.10}$$

## 7.6 Dynamic fits model 2

### 7.6.1 Peak March 1

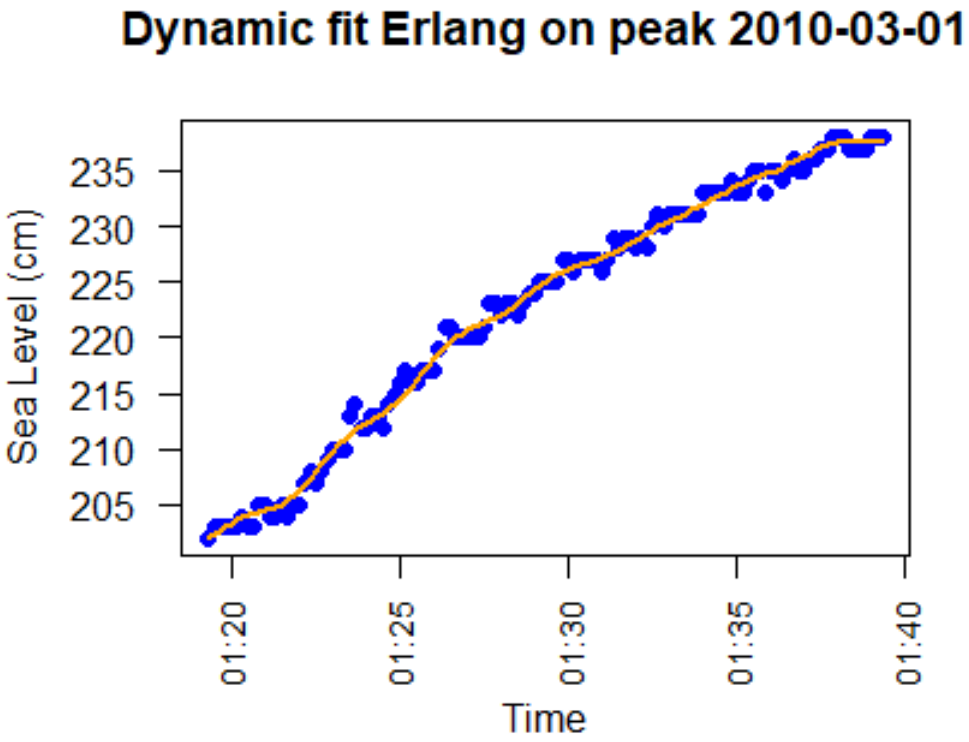


Figure 7.18: Dynamic fit second model peak March 1 10 minutes for top

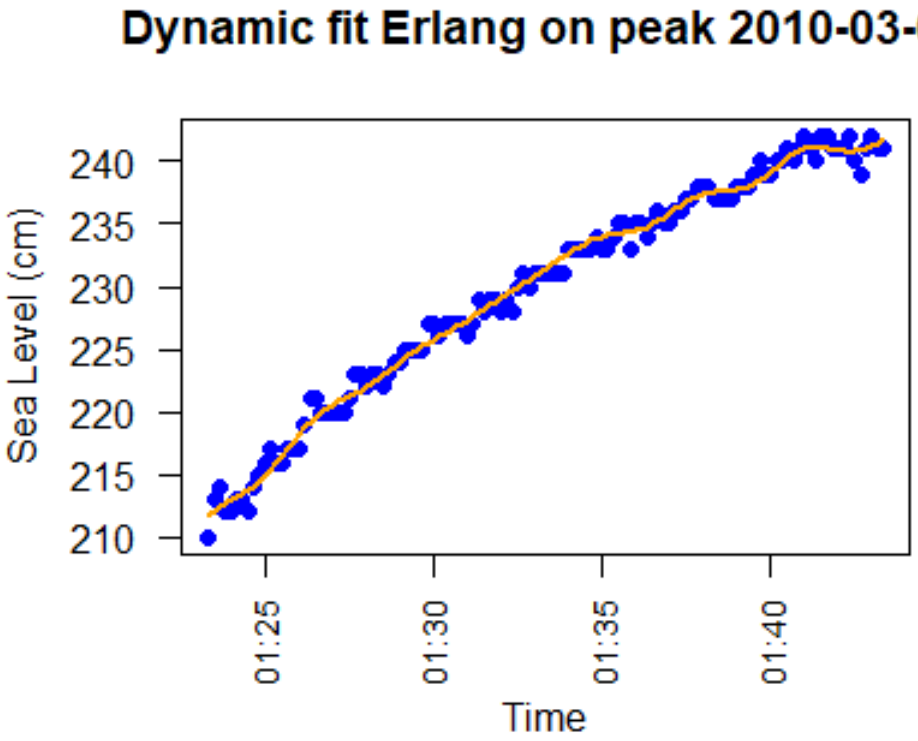


Figure 7.19: Dynamic fit second model peak March 1 5 minutes for top

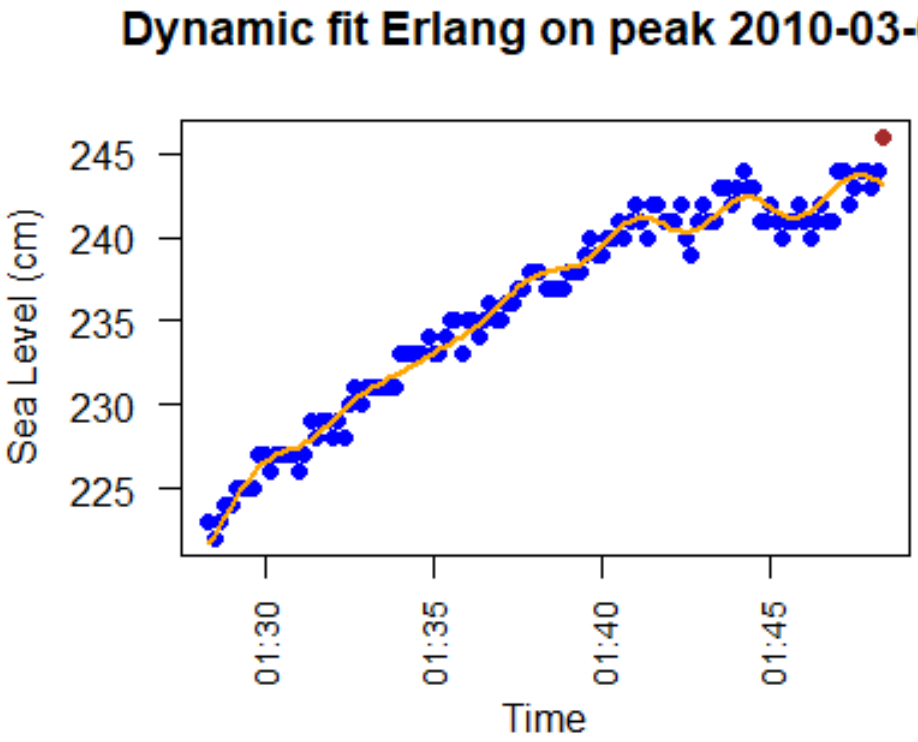


Figure 7.20: Dynamic fit second model peak March 1 0 minutes for top

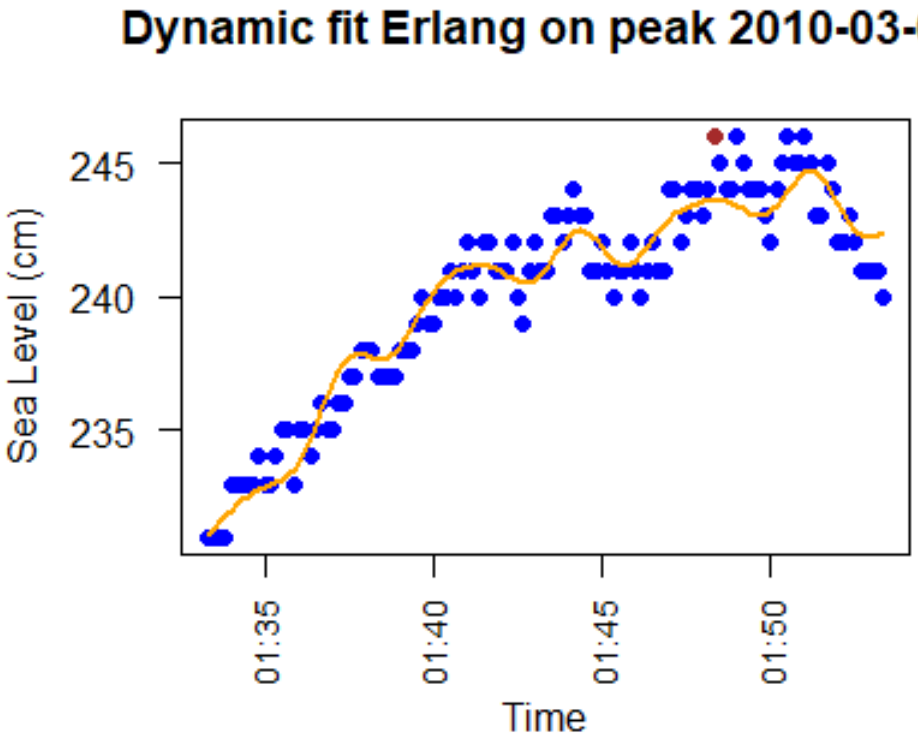


Figure 7.21: Dynamic fit second model peak March 1 5 minutes after top

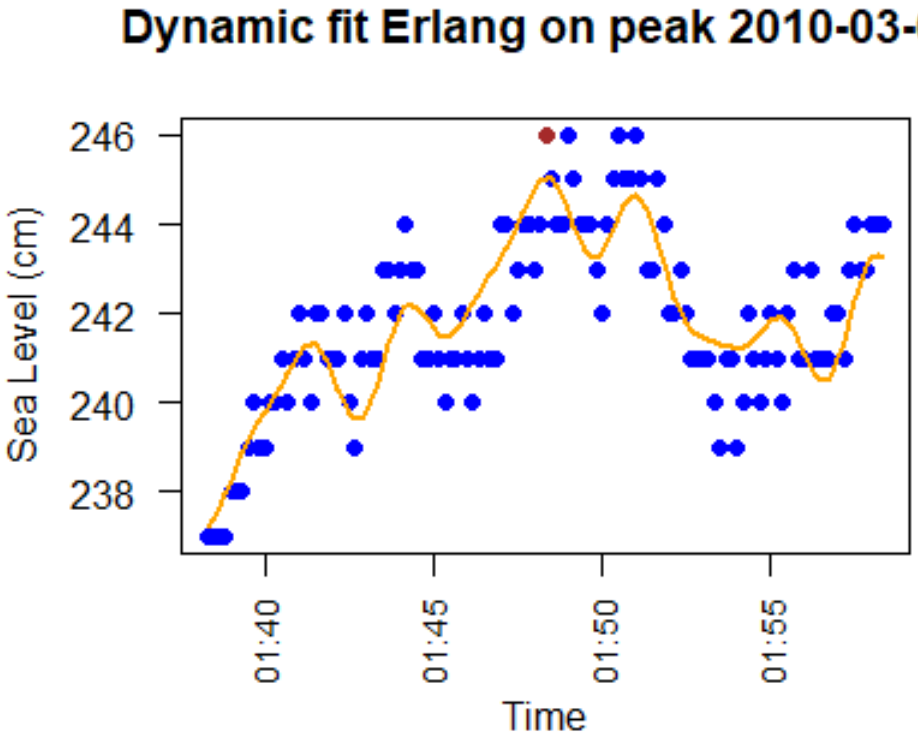


Figure 7.22: Dynamic fit second model peak March 1 10 minutes after top

| j | $endfitRng$ | $\varepsilon_{indv}^{(j)}$ |
|---|---|---|
| 1 | -600 | 0.798 |
| 2 | -540 | 0.777 |
| 3 | -480 | 0.799 |
| 4 | -420 | 0.769 |
| 5 | -360 | 0.766 |
| 6 | -300 | 0.787 |
| 7 | -240 | 0.778 |
| 8 | -180 | 0.806 |
| 9 | -120 | 0.808 |
| 10 | -60 | 0.842 |
| 11 | 0 | 0.822 |
| 12 | 60 | 0.872 |
| 13 | 120 | 0.881 |
| 14 | 180 | 0.892 |
| 15 | 240 | 0.948 |
| 16 | 300 | 1.003 |
| 17 | 360 | 1.075 |
| 18 | 420 | 1.047 |
| 19 | 480 | 1.051 |
| 20 | 540 | 1.068 |
| 21 | 600 | 1.300 |

Table 7.12: Dynamic fit of second model on peak March 1: individual error of 10 up to 10 minutes relative to the top

Combining the first layer individual errors $\varepsilon_{indv}^{(j)}$, we obtain the second layer error:

$$\varepsilon_{allavr} = 0.887 \tag{7.11}$$

### 7.6.2 Peak August 29

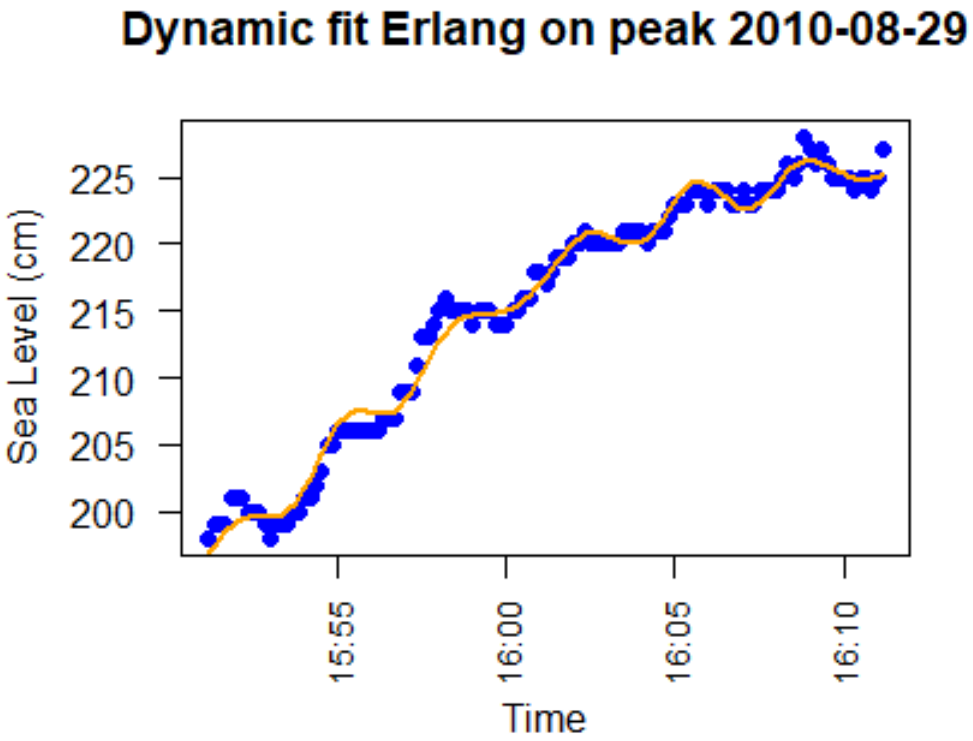


Figure 7.23: Dynamic fit second model peak August 29 10 minutes for top

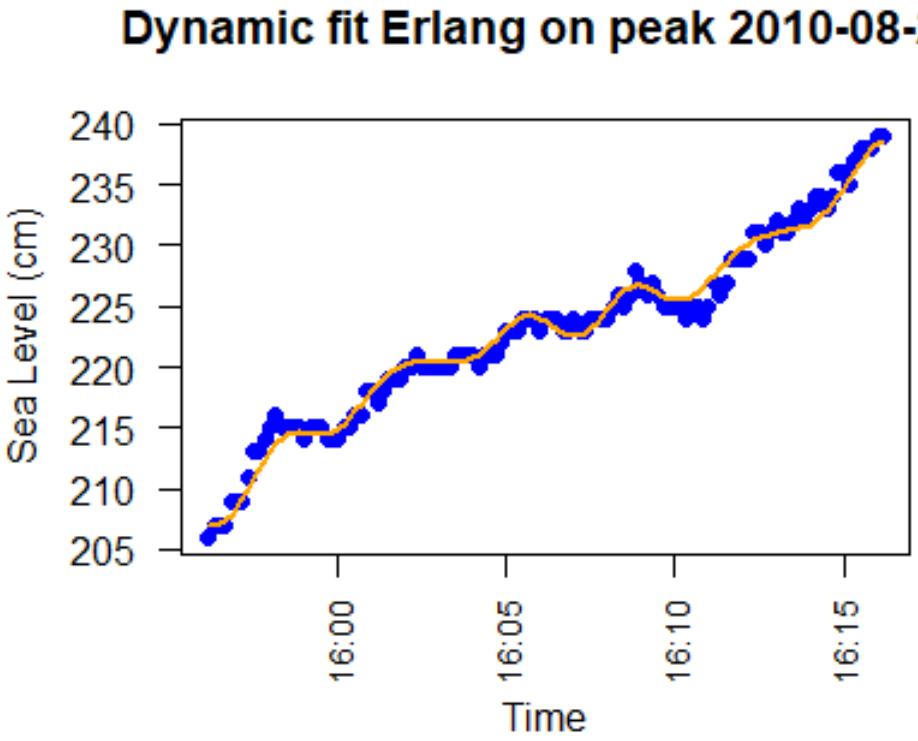


Figure 7.24: Dynamic fit second model peak August 29 5 minutes for top

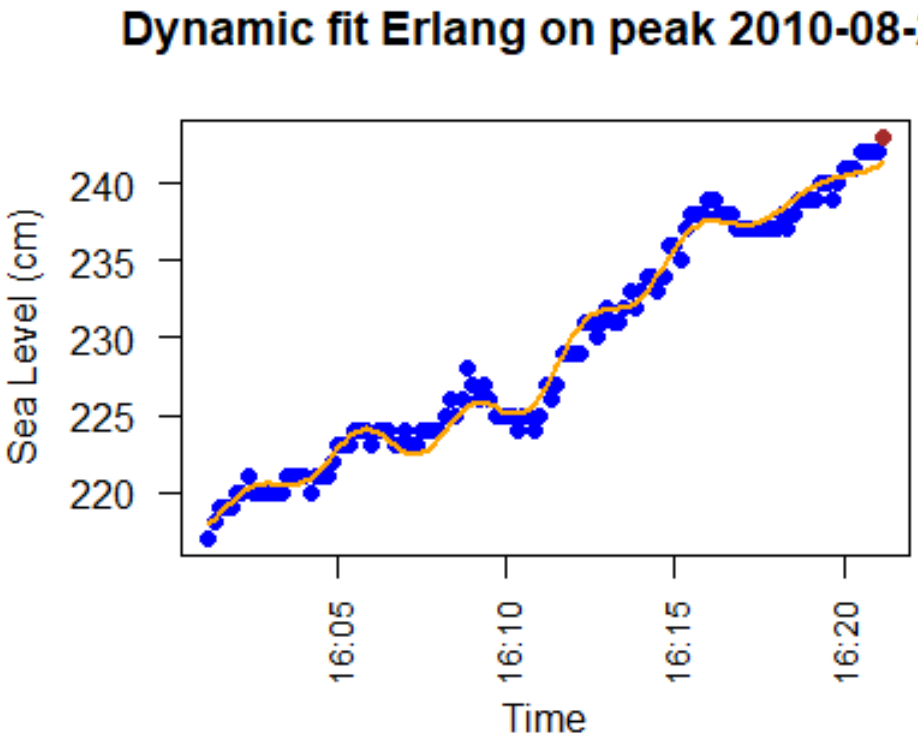


Figure 7.25: Dynamic fit second model peak August 29 0 minutes for top

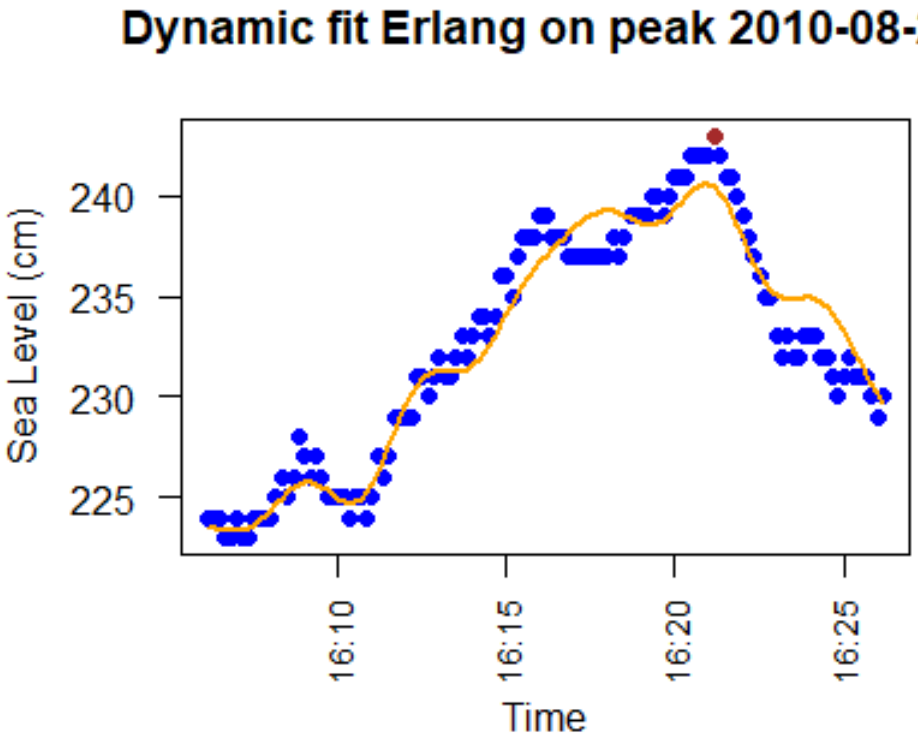


Figure 7.26: Dynamic fit second model peak August 29 5 minutes after top

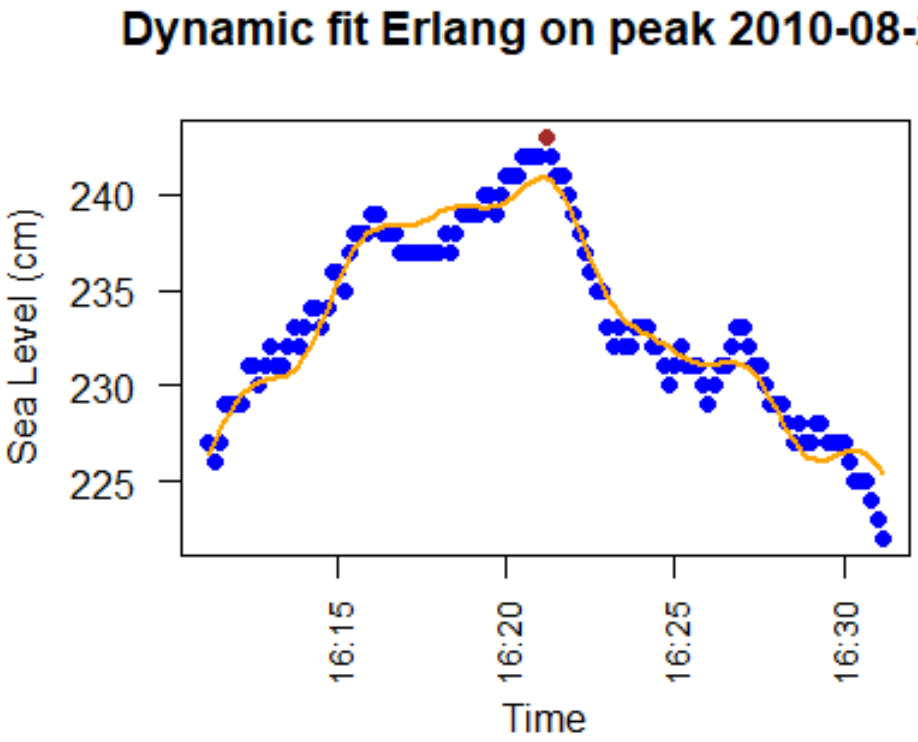


Figure 7.27: Dynamic fit second model peak August 29 10 minutes after top

| j | $endfitRng$ | $\varepsilon_{indv}^{(j)}$ |
|---|---|---|
| 1 | -600 | 0.957 |
| 2 | -540 | 0.977 |
| 3 | -480 | 0.917 |
| 4 | -420 | 1.065 |
| 5 | -360 | 1.042 |
| 6 | -300 | 0.926 |
| 7 | -240 | 1.004 |
| 8 | -180 | 0.816 |
| 9 | -120 | 0.821 |
| 10 | -60 | 0.774 |
| 11 | 0 | 0.805 |
| 12 | 60 | 0.965 |
| 13 | 120 | 1.742 |
| 14 | 180 | 1.711 |
| 15 | 240 | 1.433 |
| 16 | 300 | 1.405 |
| 17 | 360 | 1.924 |
| 18 | 420 | 1.965 |
| 19 | 480 | 1.521 |
| 20 | 540 | 1.098 |
| 21 | 600 | 1.158 |

Table 7.13: Dynamic fit of second model on peak August 29: individual error of 10 up to 10 minutes relative to the top

Combining the first layer individual errors $\varepsilon_{indv}^{(j)}$, we obtain the second layer error:

$$\varepsilon_{allavr} = 1.192 \tag{7.12}$$

### 7.6.3 Peak August 30

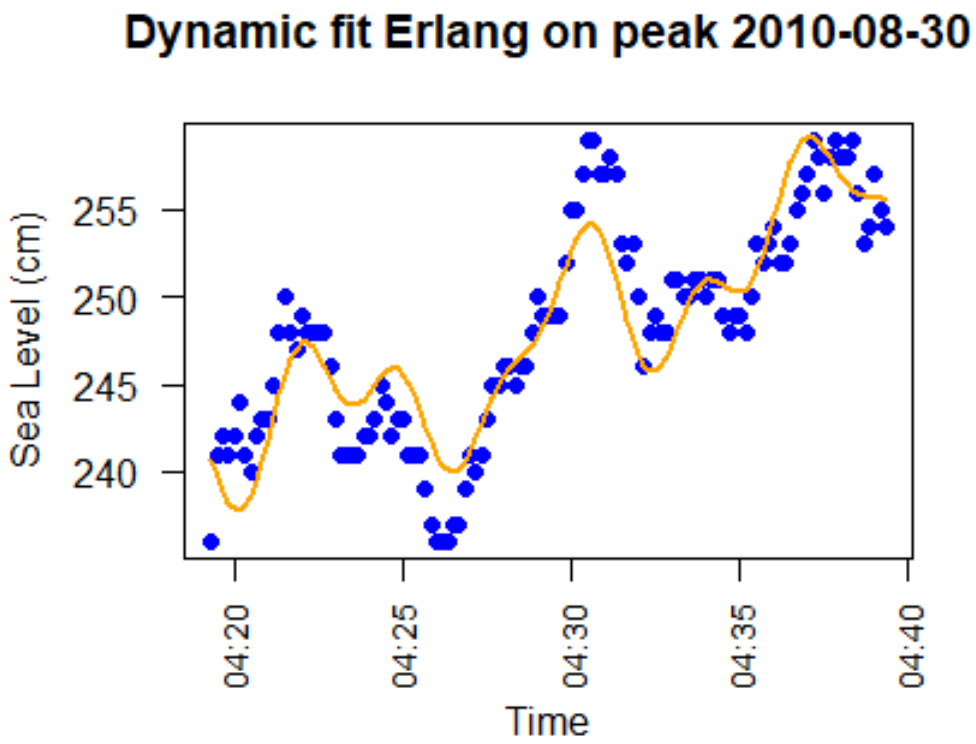


Figure 7.28: Dynamic fit second model peak August 30 10 minutes before top

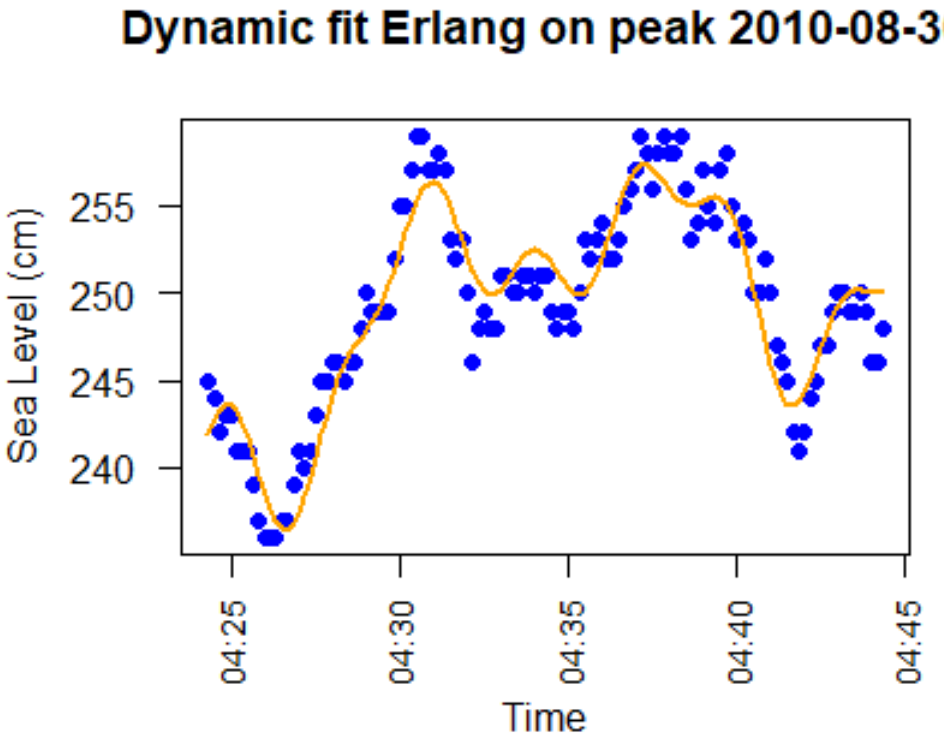


Figure 7.29: Dynamic fit second model peak August 30 5 minutes before top

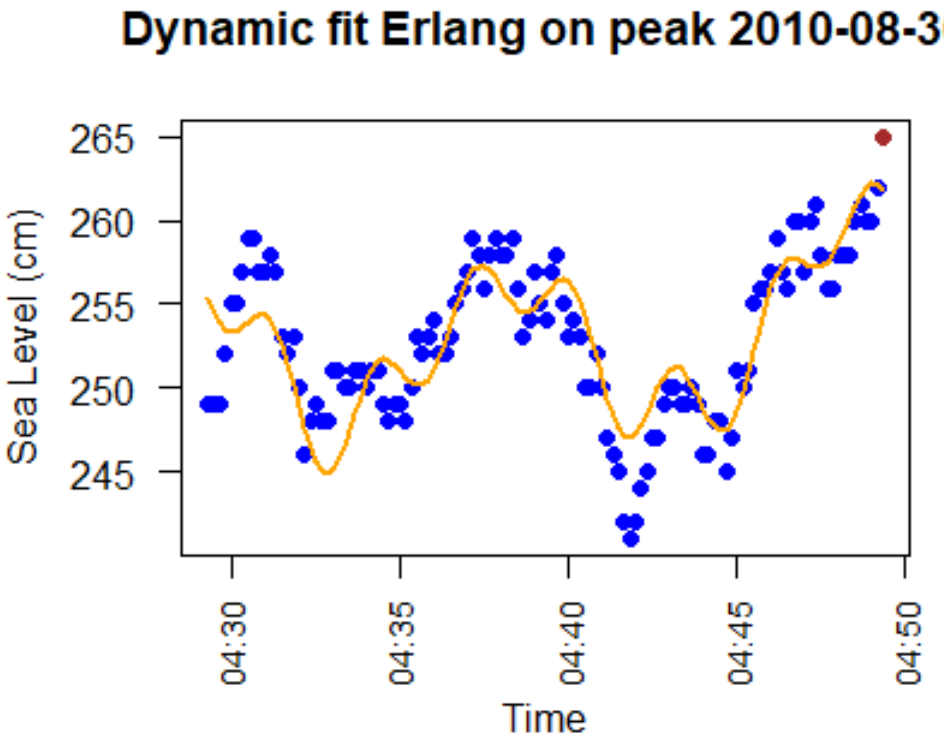


Figure 7.30: Dynamic fit second model peak August 30 0 minutes before top

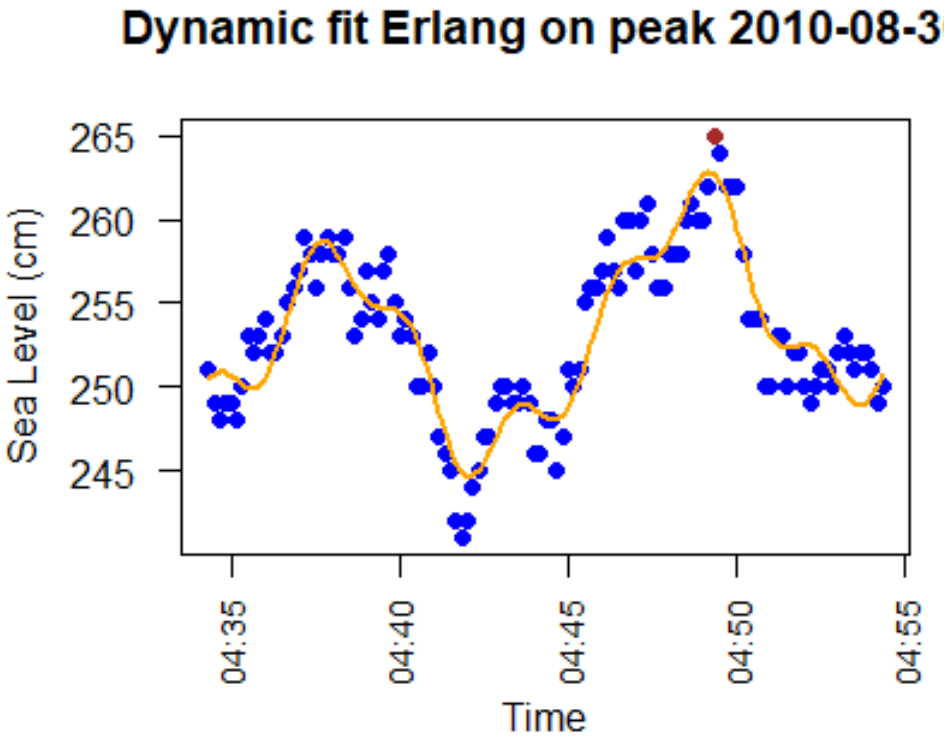


Figure 7.31: Dynamic fit second model peak August 30 5 minutes after top

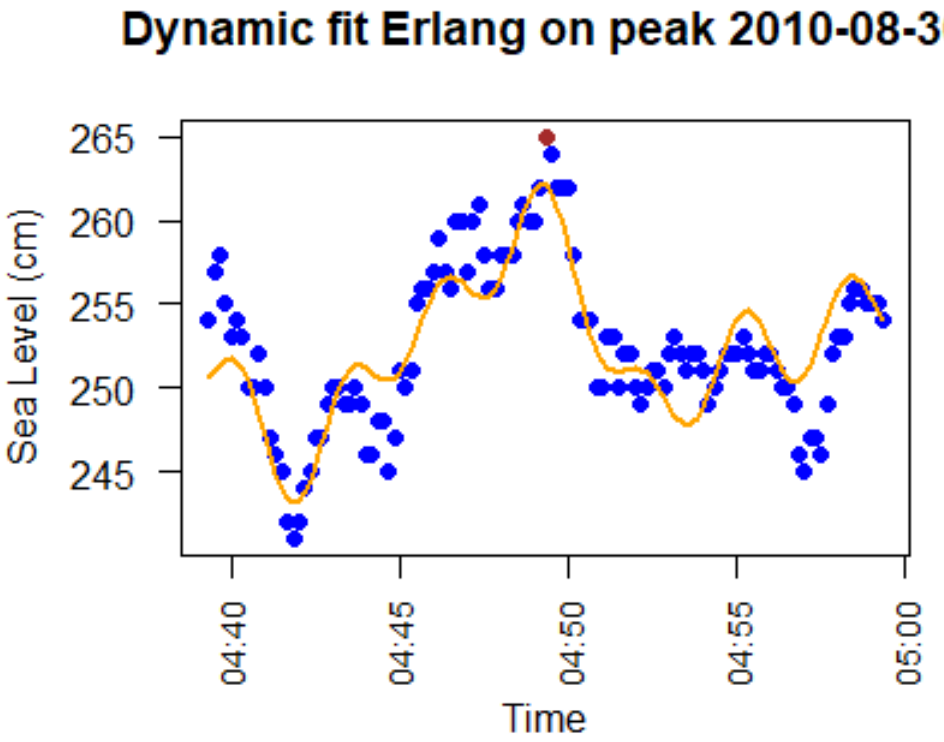


Figure 7.32: Dynamic fit second model peak August 30 10 minutes after top

| j | $endfitRng$ | $\varepsilon_{indv}^{(j)}$ |
|---|---|---|
| 1 | -600 | 2.634 |
| 2 | -540 | 2.517 |
| 3 | -480 | 2.557 |
| 4 | -420 | 2.758 |
| 5 | -360 | 1.931 |
| 6 | -300 | 1.922 |
| 7 | -240 | 2.347 |
| 8 | -180 | 2.685 |
| 9 | -120 | 3.275 |
| 10 | -60 | 3.251 |
| 11 | 0 | 2.560 |
| 12 | 60 | 2.219 |
| 13 | 120 | 2.299 |
| 14 | 180 | 2.273 |
| 15 | 240 | 2.058 |
| 16 | 300 | 1.927 |
| 17 | 360 | 2.295 |
| 18 | 420 | 2.416 |
| 19 | 480 | 2.855 |
| 20 | 540 | 2.499 |
| 21 | 600 | 2.565 |

Table 7.14: Dynamic fit of second model on peak August 30: individual error of 10 up to 10 minutes relative to the top

Combining the first layer individual errors $\varepsilon_{indv}^{(j)}$, we obtain the second layer error:

$$\varepsilon_{allavr} = 2.469 \tag{7.13}$$

## 7.7 Comparison first and second model

On March 1, the models exchange which performs better a lot of times. The invariant periods, i.e. the periods within two exchanges, is approximately 4 updates. The differences between the individual errors are quite small, they all lie within 0.25 centimeters. This is reflected in the second layer error, with $\varepsilon_{allavr} = 0.886$ for the first model and $\varepsilon_{allavr} = 0.887$ for the second model, the difference between the second layer error is only 0.001.

On August 29, the situation is different than on March 1. We see larger differences between the two models, and the large "exchanging behaviour" between the two models in the individual fits that was present on March 1 is absent here. We can safely say that the first model outperforms the second at this peak: except at the first fit (10 minutes before the top), one fit in the middle (1 minute after the top) and the last two fits (9 and 10 minutes after the top), the first model has a smaller first layer error $\varepsilon_{indv}^{(j)}$ than the second model. Although the smallest difference between these errors is still 0.004 centimeter, the largest difference now is 0.55 centimeter. If we compare the second layer errors, we continue to see the larger difference between the two models: we have $\varepsilon_{allavr} = 1.105$ for the first model and $\varepsilon_{allavr} = 1.192$ for the second. Nevertheless, the error difference is still within 0.1 centimeter.

On August 30, the models perform almost equally good, and the "exchanging behaviour" we saw earlier at the peak of March 1 we again see at this peak. The invariant periods differ a bit in length: there are periods of only 1 and 2 subsequent updates, but also instances of 4 and 5 subsequent updates. Also the differences in this "exchanging behaviour" are larger than on March 1: the largest difference on August 30 is 0.567 while the smallest is 0.06. The difference between the second layer error is intermediate: with $\varepsilon_{allavr} = 2.415$ for the first model and $\varepsilon_{allavr} = 2.469$ for the second, the first model performs slightly better, but the difference is only 0.054 centimeters.

We thus see that at March 1, the models perform almost equally good, and on August 29 and August 30, the first model performs slightly better. We expected that the second model would perform better on the peaks of March 1 and August 29, and that on August 30, the second model would not perform better (i.e. perform the same or worse). We conclude with saying that in terms of dynamic fitting, the second new model with the density of the Erlang distribution as trend does not outperform the existing, first model with the parabolic trend.

## 7.8 Conclusion

Different than with static fitting, in dynamic fitting we do not know the location of the top. To get an idea of the location, we have proposed various estimators. In the end, we decided that the most suitable estimator is the so-called "moving minimum" estimator, formally defined as

$$\widetilde{loctop}_{peakNr}^{(endfitRng)} := \{\widehat{loctop}_{peakNr}^{(endfitRng)} | \min_{endfitRng \in \{allendfitRng\}} (RSS(y_i^{(endfitRng)}))\} \tag{7.14}$$

The main advantage of this estimator over the other ones is that it is able to filter out the estimates with a relatively high bias (of more than 40 minutes).
Next to the explicit parameters in the model, the size of the fit range is an implicit parameter. To find the optimal size, we dynamically fitted both the first and the second model with 900, 1200, 1500 and 1800 seconds for multiple fits. For all fits of both models a fit range of 900 seconds leads to the lowest layer one error $\varepsilon_{indv}^{(j)}$. However, since we want the fit range to include 2 periods of the short-term oscillations ($= 2 \cdot 545 = 1090$ seconds), we choose a fit range of 1200 seconds for the predictions.
With this fit range, we dynamically fitted both models. The models perform differently over the various peaks: at March 1 they perform almost equally good and on August 29 and 30, the first model performs slightly better. Thus, in terms of dynamic fitting, the second model, with the density of the Erlang distribution as trend, is not able to outperform the existing, first model with the parabola as trend.

# Chapter 8

# Model predictions

In this chapter, we investigate the performance of the model, or more specifically, we investigate the performance of its predictions. We first investigated the optimal fit range by performing predictions of all peaks on the first model and calculating the total error $\varepsilon$, and doing so for multiple fit ranges. We continued with the fit range with the lowest error $\varepsilon$, 35 minutes. The model (first or second) is fitted 21 times, where the end of the fit range (and thus the beginning of the prediction range) is positioned at 10 minutes before the top up to 10 minutes after the top. Indeed, the step size for an update is 1 minute, in line with the requirement to make a decision every minute. Based on each fit, a 5 minute-ahead prediction is made. For each prediction, the first layer error $\varepsilon_{indv}^{(j)}$ is calculated and tabulated. The eighth fit and prediction, that is the one with $endfitRng = 3$ minutes before the top is plotted, where the actual water level is included for comparison. After all peaks, the overall, second layer errors $\varepsilon_{allavr}$ are tabulated and the total error $\varepsilon$ is calculated.

## 8.1 Predictions first model

### 8.1.1 Experimenting with the fitting range

Just like with dynamic fitting, the size of the fitting range is another parameter that we can choose. In dynamic fitting the fitting range leading to the lowest error was the lowest fitting range. This does, however, not be the case with the predictions. To experiment we have taken fit ranges from 20 minutes up to 45 minutes with intermediate step size of 5 minutes, and computed the total, layer three error $\varepsilon$, using the three peaks we also considered with the dynamic fitting in Chapter 7: March 1, August 29 and August 30.

| $fitRng$ | $\varepsilon$ |
|---|---|
| 20 | 7.298 |
| 25 | 5.363 |
| 30 | 5.396 |
| 35 | 5.170 |
| 40 | 5.263 |
| 45 | 5.386 |

Table 8.1: Total error for multiple fit ranges of the first model

As we can see in Table 8.1, for the first model the lowest total error over the peaks occurs at a fit range of 35 minutes. Hence we will continue our predictions for the first model with a fit range of 35 minutes.

### 8.1.2 Predictions peak March 1

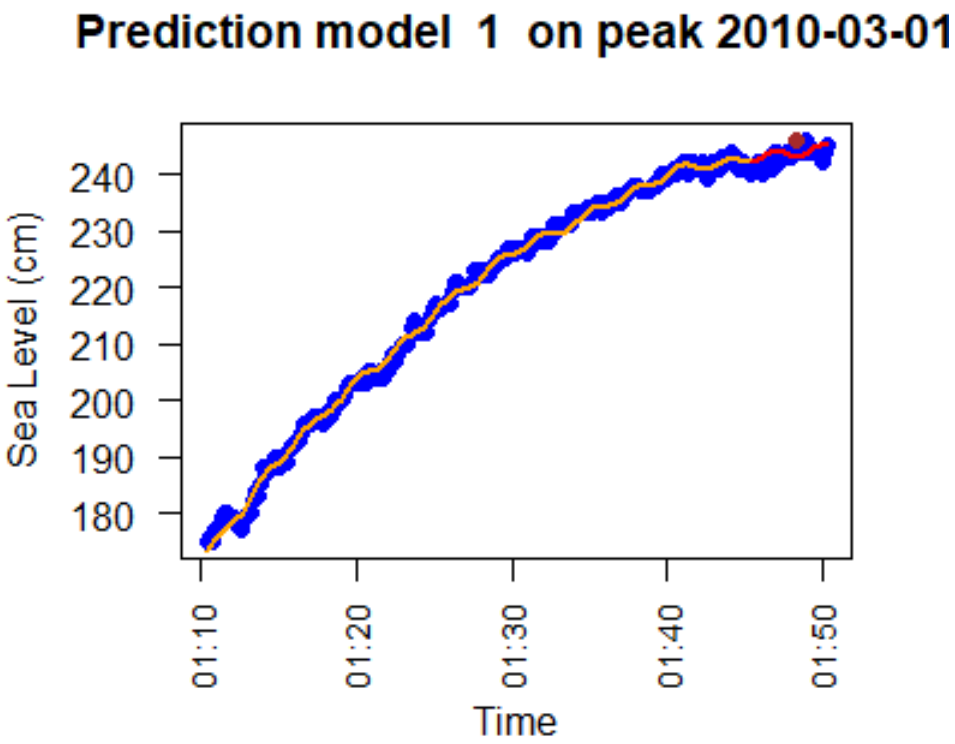


Figure 8.1: Prediction first model peak March 1 with $endfitRng$ 3 minutes for top

| j | $endfitRng$ | $\varepsilon_{indv}^{(j)}$ |
|---|---|---|
| 1 | -600 | 2.187 |
| 2 | -540 | 1.695 |
| 3 | -480 | 1.924 |
| 4 | -420 | 2.668 |
| 5 | -360 | 2.981 |
| 6 | -300 | 1.661 |
| 7 | -240 | 1.789 |
| 8 | -180 | 1.693 |
| 9 | -120 | 1.417 |
| 10 | -60 | 1.662 |
| 11 | 0 | 1.973 |
| 12 | 60 | 2.624 |
| 13 | 120 | 2.553 |
| 14 | 180 | 2.749 |
| 15 | 240 | 2.247 |
| 16 | 300 | 2.792 |
| 17 | 360 | 3.816 |
| 18 | 420 | 4.798 |
| 19 | 480 | 4.257 |
| 20 | 540 | 3.929 |
| 21 | 600 | 2.566 |

Table 8.2: Peak March 1: individual error of predictions 10 minutes before up to 10 minutes after the top

Taking the average over all these first layer errors, we obtain the second layer error:

$$\varepsilon_{allavr} = 2.571 \tag{8.1}$$

### 8.1.3 Predictions peak August 29

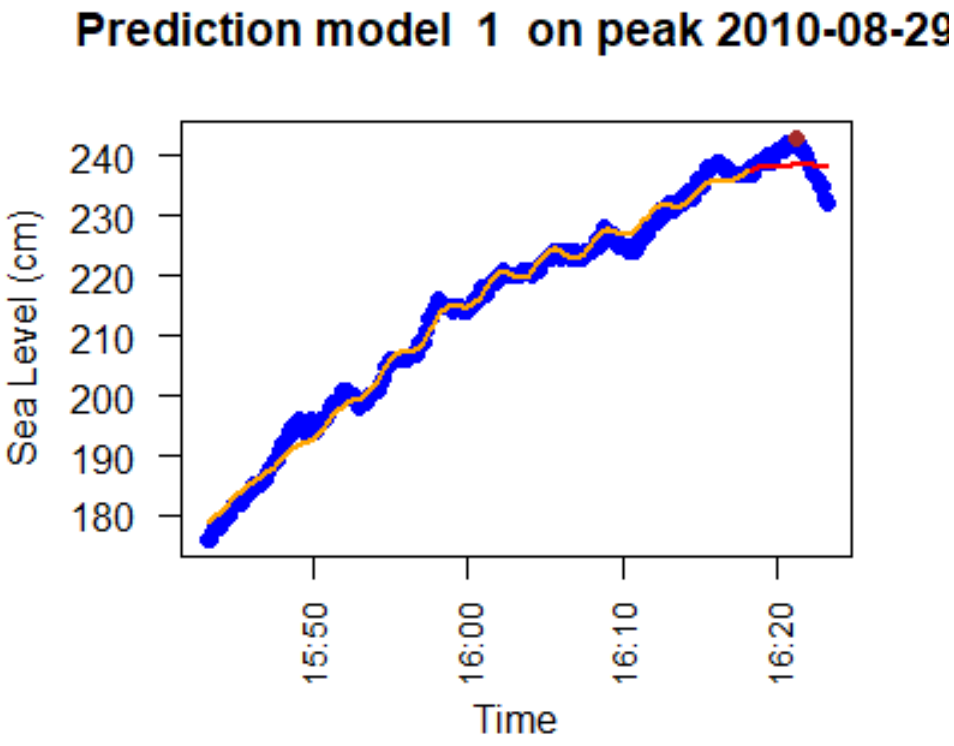


Figure 8.2: Prediction first model peak August 29 with *endfitRng* 3 minutes for top

| j | $endfitRng$ | $\varepsilon_{indv}^{(j)}$ |
|---|---|---|
| 1 | -600 | 3.092 |
| 2 | -540 | 5.172 |
| 3 | -480 | 5.840 |
| 4 | -420 | 6.557 |
| 5 | -360 | 6.796 |
| 6 | -300 | 5.288 |
| 7 | -240 | 4.024 |
| 8 | -180 | 2.724 |
| 9 | -120 | 3.906 |
| 10 | -60 | 5.645 |
| 11 | 0 | 7.593 |
| 12 | 60 | 8.894 |
| 13 | 120 | 8.777 |
| 14 | 180 | 7.339 |
| 15 | 240 | 5.854 |
| 16 | 300 | 5.702 |
| 17 | 360 | 8.327 |
| 18 | 420 | 9.864 |
| 19 | 480 | 10.836 |
| 20 | 540 | 11.716 |
| 21 | 600 | 10.917 |

Table 8.3: Peak August 29: individual error of predictions 10 minutes before up to 10 minutes after the top

Taking the average over all these first layer errors, we obtain the second layer error:

$$\varepsilon_{allavr} = 6.898 \tag{8.2}$$

### 8.1.4 Predictions peak August 30

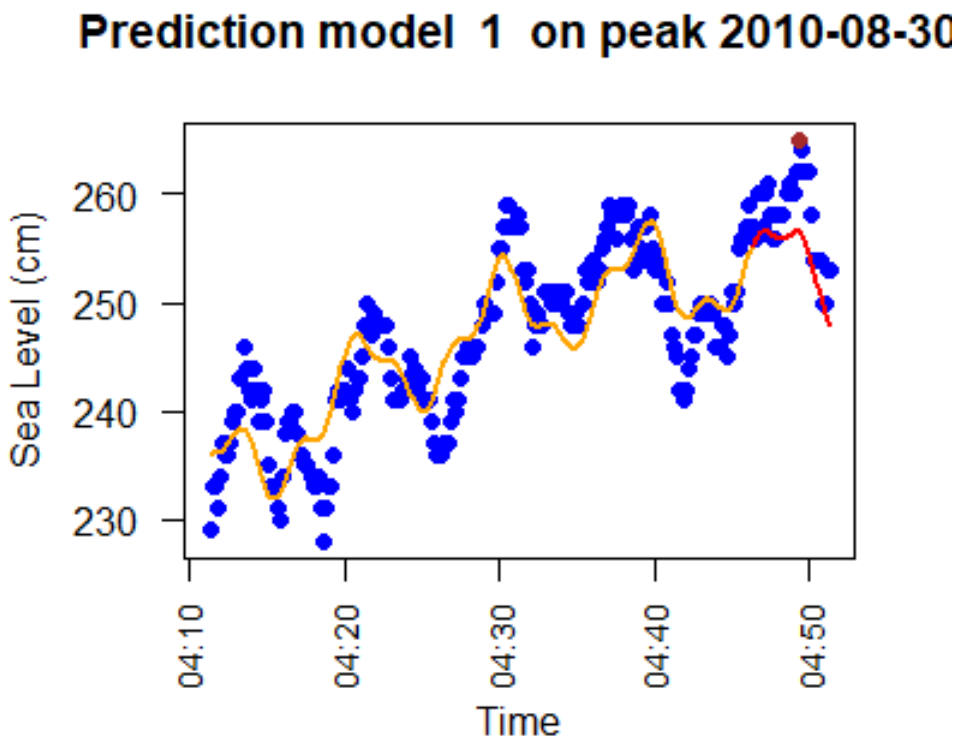


Figure 8.3: Prediction first model peak August 30 with $endfitRng$ 3 minutes for top

| j | $endfitRng$ | $\varepsilon_{indv}^{(j)}$ |
|---|---|---|
| 1 | -600 | 10.528 |
| 2 | -540 | 9.640 |
| 3 | -480 | 6.430 |
| 4 | -420 | 4.598 |
| 5 | -360 | 4.848 |
| 6 | -300 | 5.912 |
| 7 | -240 | 6.646 |
| 8 | -180 | 4.112 |
| 9 | -120 | 3.365 |
| 10 | -60 | 3.233 |
| 11 | 0 | 3.321 |
| 12 | 60 | 2.543 |
| 13 | 120 | 4.076 |
| 14 | 180 | 5.325 |
| 15 | 240 | 6.377 |
| 16 | 300 | 7.088 |
| 17 | 360 | 6.791 |
| 18 | 420 | 6.337 |
| 19 | 480 | 7.741 |
| 20 | 540 | 9.797 |
| 21 | 600 | 8.132 |

Table 8.4: Peak August 30: individual error of predictions 10 minutes before up to 10 minutes after the top

Taking the average over all these first layer errors, we obtain the second layer error:

$$\varepsilon_{allavr} = 6.040 \tag{8.3}$$

To compare between the peaks, we tabulated the second layer error amongst the three peaks

| date | $nrPeak$ | $\varepsilon_{allavr}$ |
|---|---|---|
| March 1 | 3 | 2.571 |
| August 29 | 5 | 6.898 |
| August 30 | 1 | 6.040 |

Table 8.5: Average overall error per peak

Averaging over the overall errors of all peaks, we see that the total error is

$$\varepsilon = 5.170 \tag{8.4}$$

## 8.2 Predictions second model

### 8.2.1 Predictions March 1

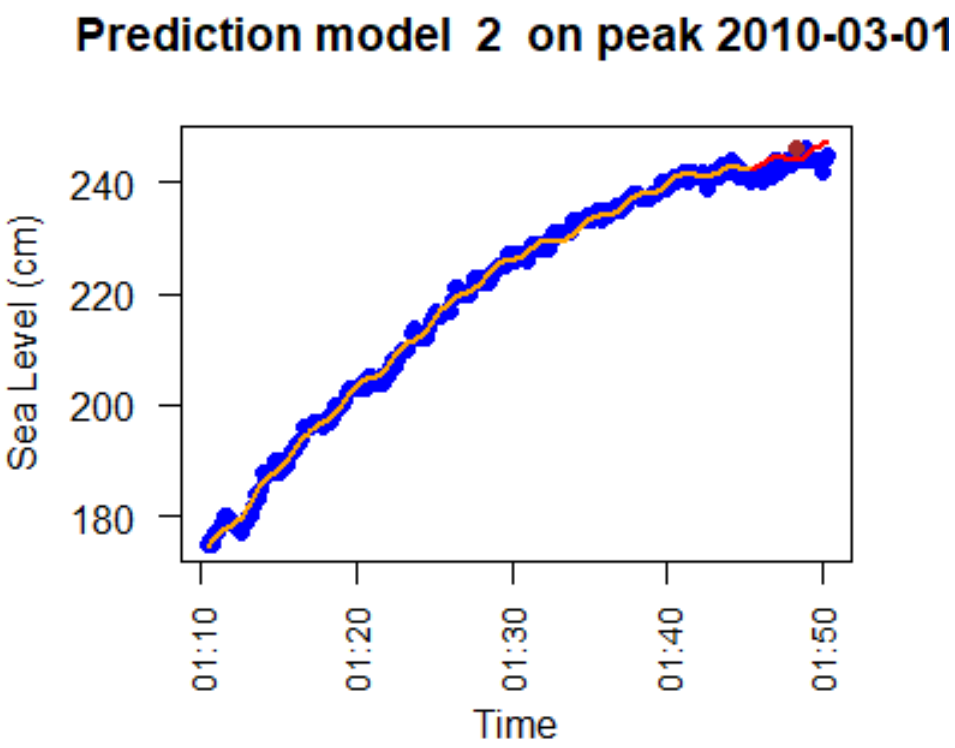


Figure 8.4: Prediction second model peak March 1 with *endfitRng* 3 minutes for top

| j | $endfitRng$ | $\varepsilon_{indv}^{(j)}$ |
|---|---|---|
| 1 | -600 | 1.839 |
| 2 | -540 | 1.342 |
| 3 | -480 | 1.594 |
| 4 | -420 | 1.541 |
| 5 | -360 | 1.996 |
| 6 | -300 | 1.728 |
| 7 | -240 | 1.815 |
| 8 | -180 | 2.070 |
| 9 | -120 | 1.834 |
| 10 | -60 | 2.176 |
| 11 | 0 | 2.737 |
| 12 | 60 | 3.617 |
| 13 | 120 | 4.135 |
| 14 | 180 | 4.386 |
| 15 | 240 | 4.250 |
| 16 | 300 | 3.098 |
| 17 | 360 | 1.774 |
| 18 | 420 | 2.110 |
| 19 | 480 | 2.589 |
| 20 | 540 | 1.986 |
| 21 | 600 | 1.997 |

Table 8.6: Peak March 1: individual error of predictions 10 minutes before up to 10 minutes after the top

Taking the average over all these first layer errors, we obtain the second layer error:

$$\varepsilon_{allavr} = 2.410 \tag{8.5}$$

### 8.2.2 Predictions peak August 29

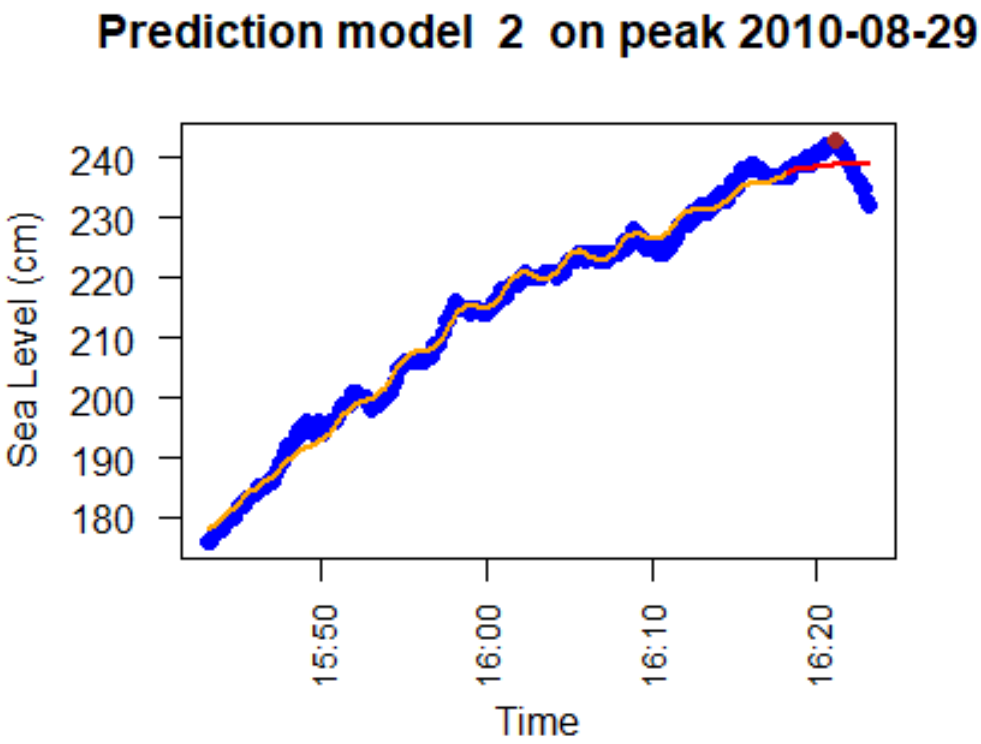


Figure 8.5: Prediction second model peak August 29 with $endfitRng$ 3 minutes for top

| j | $endfitRng$ | $\varepsilon_{indv}^{(j)}$ |
|---|---|---|
| 1 | -600 | 5.892 |
| 2 | -540 | 5.138 |
| 3 | -480 | 4.927 |
| 4 | -420 | 4.054 |
| 5 | -360 | 3.314 |
| 6 | -300 | 2.471 |
| 7 | -240 | 2.044 |
| 8 | -180 | 2.757 |
| 9 | -120 | 4.011 |
| 10 | -60 | 5.350 |
| 11 | 0 | 7.845 |
| 12 | 60 | 8.682 |
| 13 | 120 | 7.979 |
| 14 | 180 | 8.175 |
| 15 | 240 | 7.218 |
| 16 | 300 | 6.931 |
| 17 | 360 | 8.726 |
| 18 | 420 | 11.013 |
| 19 | 480 | 11.082 |
| 20 | 540 | 11.649 |
| 21 | 600 | 10.646 |

Table 8.7: Peak August 29: individual error of predictions 10 minutes before up to 10 minutes after the top

Taking the average over all these first layer errors, we obtain the second layer error:

$$\varepsilon_{allavr} = 6.662 \tag{8.6}$$

### 8.2.3 Predictions peak August 30

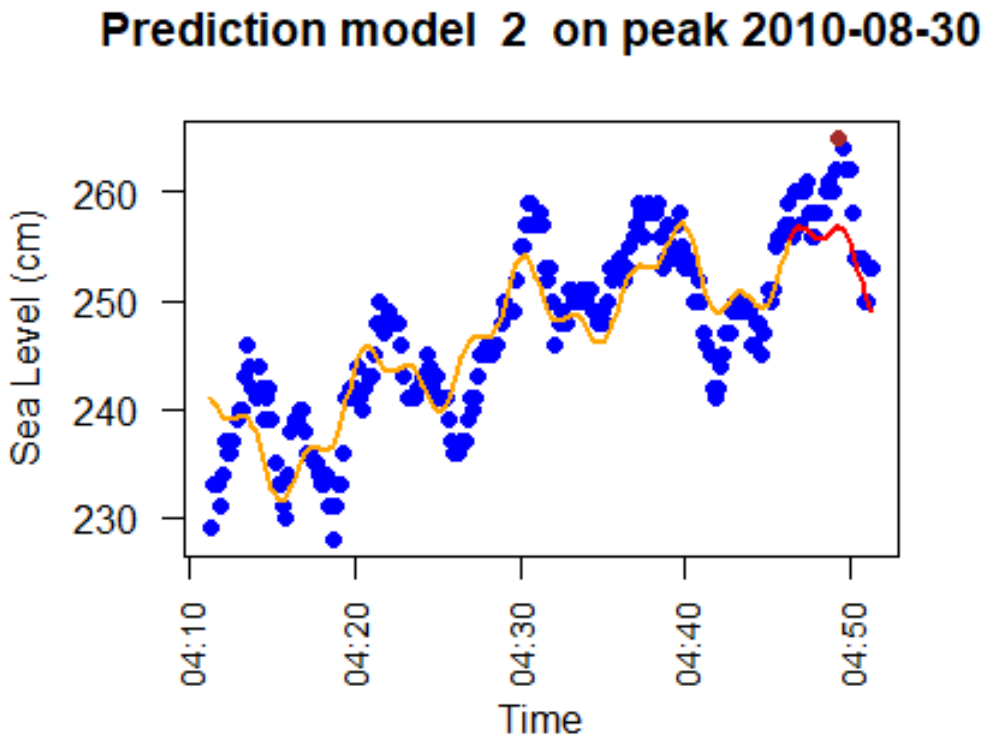


Figure 8.6: Prediction second model peak August 30 with *endfitRng* 3 minutes for top

| j | $endfitRng$ | $\varepsilon_{indv}^{(j)}$ |
|---|---|---|
| 1 | -600 | 10.007 |
| 2 | -540 | 9.571 |
| 3 | -480 | 6.007 |
| 4 | -420 | 3.859 |
| 5 | -360 | 3.587 |
| 6 | -300 | 4.383 |
| 7 | -240 | 4.867 |
| 8 | -180 | 3.826 |
| 9 | -120 | 3.689 |
| 10 | -60 | 3.363 |
| 11 | 0 | 3.477 |
| 12 | 60 | 3.143 |
| 13 | 120 | 4.110 |
| 14 | 180 | 6.051 |
| 15 | 240 | 7.353 |
| 16 | 300 | 7.716 |
| 17 | 360 | 7.369 |
| 18 | 420 | 6.421 |
| 19 | 480 | 6.092 |
| 20 | 540 | 6.182 |
| 21 | 600 | 5.915 |

Table 8.8: Peak August 29: individual error of predictions 10 minutes before up to 10 minutes after the top

Taking the average over all these first layer errors, we obtain the second layer error:

$$\varepsilon_{allavr} = 5.571 \tag{8.7}$$

To compare between the peaks, we tabulated the second layer error amongst the three peaks

| date | $nrPeak$ | $\varepsilon_{allavr}$ |
|---|---|---|
| March 1 | 3 | 2.410 |
| August 29 | 5 | 6.662 |
| August 30 | 1 | 5.571 |

Table 8.9: Average overall error per peak

Averaging over the overall errors of all peaks, we see that the total error is

$$\varepsilon = 4.881 \tag{8.8}$$

## 8.3 Comparing between predictions first model and second model

On March 1, each of the two models performs best on some part of the peak, so that there is no overall best model. There are only two switches between the two models. The differences between the two models can be quite large but also quite small: the largest difference is 2.668 centimeter, while the smallest difference is 0.026 centimeter. A difference in overall performance does become visible if we consider the second layer error: with $\varepsilon_{allavr} = 2.571$ centimeter for the first model and $\varepsilon_{allavr} = 2.410$ centimeter for the second model, the second model performs better on March 1 in terms of the overall error.

On August 29, the models exchange which performs better a lot of time: in total, there are seven exchanges. At this peak, the difference in error between the two models along the different updates varies even more than on March 1: the largest difference is 3.482, while the smallest difference is 0.033 centimeter. The individual errors are quite high in comparison with the first peak on March 1: for both models, the highest error is more than 11 centimeters. If we consider the second layer error, we can see that the overall performance of the two models stays close: with $\varepsilon_{allavr} = 6.898$ for the first model and $\varepsilon_{allavr} = 6.662$ for the second model, the difference is only 0.236 centimeters.

On August 30, the situation is again similar to that on March 1: none of the two models performs best on all updates of the fits, but there are only two exchanges in best model, and the invariant periods (where one model performs best) are quite long. Again, there is quite a large range in the differences of the errors of the models: the smallest difference is 0.034 centimeter, whereas the largest difference is 3.615 centimeter. Similar to August 29, the errors can become quite large: also at this peak there are occurrences of an error larger than 10 centimeters at both models. Considering the second layer error, we see that the difference in overall performance of the peak is somewhat higher for this peak than for August 29 and March 1: with $\varepsilon_{allavr} = 6.040$ for the first model and $\varepsilon_{allavr} = 5.571$ for the second model, for this peak the difference is 0.469 centimeter.

Considering the overall, second layer error, we thus see that the second model outperforms the first model on all considered peaks. For the peaks of March 1 and August 29 this was within our expectation or at least our hope, as these were the peaks were the second model was especially designed for. However, the "benchmark" peak of August 30 shows that the second model can also be a better alternative for other peaks.

## 8.4 Conclusion

After experimenting with fit ranges of different sizes, it turns out that a fit range of 35 minutes leads to the lowest total error $\varepsilon$ in the predictions of the first model.
On the level of individual predictions of one peak, there is no best model: at all peaks both models have a part of the peak where they have the lowest first layer error $\varepsilon^{(j)}_{indv}$. However, on the level of the peak itself, so in terms of the second layer error $\varepsilon_{allavr}$, the second model performs better than the first model at all the three considered peaks, even at the "benchmark" peak of August 30. Hence the second model also outperforms the first in terms of the third layer error $\varepsilon$. This shows that, based on the considered peaks, the second model can be best used to predict the water level.

# Chapter 9

# Conclusions

First, we discuss the graphical description of the peaks. Considering plots of the full peaks, generally, the peaks seem to resemble a parabolic trend. However, the peaks are asymmetrical around the top: the top is skewed to the left. After a quick rise of the water level follows a slower, longer spread fall. If we zoom in on the top, we can see that most peaks include some local maxima, before the global maximum (the top) is reached. The number of these local maxima differs per peak. We come to the conclusion that the behaviour per peak differs significantly, and that for some peaks, the parabolic trend does not suffice. For these peaks, another description of the trend is necessary.

We build up three models, two of which we have investigated in further detail and compared with one another. The models differ in the way they model the trend, the short-term oscillations are constantly modelled in the same manner. The first model takes the form

$$Y_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + a_1 \sin(\omega_1 t_i + \rho_1) + a_2 \sin(\omega_2 t_i + \rho_2) + a_3 \sin(\omega_3 t_i + \rho_3) + \varepsilon_i \tag{9.1}$$

The second model takes the form

$$Y_i = \beta_0 + \lambda^2 \left( \frac{t_i}{B} + \tilde{A} \right) e^{-\lambda\left(\frac{t_i}{B} + \tilde{A}\right)} + a_1 \sin(\omega_1 t_i + \rho_1) + a_2 \sin(\omega_2 t_i + \rho_2) + a_3 \sin(\omega_3 t_i + \rho_3) + \varepsilon_i \tag{9.2}$$

In terms of static fitting, the first model has a lower first layer error $\varepsilon_{indv}^{(j)}$ than the second model for all considered peaks (March 1, August 29 and August 30). However, for the peaks that were the inspiration for the second model, March 1 and August 29, the difference in error between the two models is only 0.08 centimeter for March 1 and 0.42 centimeter for August 29. We conclude that the second model can (almost) equally well statically model these peaks as the first model, but not better. On the peak of August 30, the peak meant as benchmark, the second model performs worse than the first model, with more than 1 centimeter difference.

In terms of dynamic fitting, the models perform differently over the considered peaks. On March 1, the models perform almost equally good (the difference in the second layer error $\varepsilon_{allavr}$ is only 0.001 centimeter). On August 29 and August 30, the first model performs slightly better: the difference in error is 0.087 centimeter on August 29 and 0.054 centimeter on August 30. Thus, in terms of dynamic fitting, the second model, with the density of the Erlang distribution as trend, is not able to outperform the existing, first model with the parabola as trend.

In terms of the predictions, there is a difference in performance on the different layers. On the first layer, thus considering the individual predictions of one peak, there is no best model: at all peaks both models have a part of the peak where they have the lowest first layer error $\varepsilon_{indv}^{(j)}$. On the second layer, thus considering the full peak, the second model performs better than the first model at all the three considered peaks, even at the "benchmark" peak of August 30. We conclude to say that the second model can be best used to predict the water level.

## 9.1 Answering of the research questions

To answer (Q1) first, we captured the short-term oscillations by the form $a_i \sin(\omega_i t + \rho_i)$. We obtained the frequencies $\omega_i$ by doing a Fourier analysis on the detrended data (see Section 6.4.5) The three arguments of $\omega_j$ that yield the three highest values in the periodogram $I(\omega_j)$ were our initial "guesses" for the frequency. For the phases $\rho_i$, we tried all the of the five values $\{0, \frac{2\pi}{5}, \frac{4\pi}{5}, \frac{6\pi}{5}, \frac{8\pi}{5}\}$ as initial starting value. The amplitudes $a_i$ do not require a starting value. We found the precise values of $a_i$, $\rho_i$ and $\omega_i$ when we fitted the total model, including both the short-term oscillations and the tide. The methodology is adaptive, because every next moment we need new predictions (every minute), we detrend the data again, giving us new initial guesses for the frequencies.

While the initial guesses of the phases and amplitudes do not change over time, their actual value can change as we fitted the total model over different measurements.

Subsequently, we answer sub-research question (Q2): we have presented two models to capture the trend of the water level. The first model models the trend as a parabola. In a formula, this takes the form $\beta_0+\beta_1 t+\beta_2 t^2$. Since the parameters $\beta_j$ appear linear in the model, we don't need starting values for them. Again, the actual values for the parameters $\beta_j$ can change when the model is refitted, making the methodology adaptive. The second model models the trend as the density of the Erlang distribution. In a formula, this takes the form $\beta_0 + \lambda^2 \left(\frac{t_i}{B} + \tilde{A}\right) e^{-\lambda\left(\frac{t_i}{B}+\tilde{A}\right)}$. All parameters need starting values. For all parameters we have supplied a range of starting values, together forming a grid (in static fitting, the parameter $\lambda$ can be expressed in the other parameters, see Section 6.3). Again, the actual values for the parameters change when the model is refitted, making also this methodology adaptive.

Considering the requirements, requirements [R1] and [R2] are for filled easily; requirement [R3] is the most relevant. With the current methodology and implementation in *R*, the first model is able to deliver predictions within one minute. The second model, however, takes approximately 4 minutes to finish.

Finally, we can answer the main research question. With the proposed two models, we are not able to make a reliable 5 minute ahead prediction. The total error $\varepsilon$ that both models make in the predictions is approximately 4 centimeters higher than the maximal allowed (by the Dutch Ministry of Water Management and Infrastructure) error $\varepsilon = 1$. Nevertheless, the second model is an improvement on the existing, first model. The first model can predict the water level with $\varepsilon = 5.170$ centimeter, while the second model can predict the water level with $\varepsilon = 4.881$ centimeter. Thus, if we needed to choose between the two models, we would go for the second model.

# Chapter 10

# Possible extensions of the project and future research

- For the first model, we established practical estimations of the theoretical boundary points for the model, i.e. the inflection points of the measurements. For the left boundary point, we did this by first finding the global minimum and then approaching the boundary point from below. The location we determined for this global minimum is relatively rough. We found the values of the parameters in the procedure by some manual experimenting. Instead, we could implement a more systematic approach that lead to a better estimate of the global minimum. This could be done by writing a program that computes, for a large sequence of vectors of parameter values, the sum of the differences between the actual and the estimated global minima amongst all peaks. In that manner, we can choose the vector of parameter values that lead to the smallest sum of differences, and thus the best estimation of the global minima. We can then determine the actual global minimum by another, much more reliable method.

- In detrending the data, we developed a performance measure that splits the detrended data in at most $k$ pieces, where $k = \lceil \frac{fitrange}{1500} \rceil$, i.e. the nearest integer greater than $\frac{fitrange}{1500}$, and then we have taken the average value over that piece. However, due to oscillations, it sometimes happens that there almost only high positive, or low negative values, so that the absolute value of the average could give a high value, while the trend is captured adequately. To overcome this problem, we could first fit a high degree polynomial over the data, before we calculate the averages. More precisely, we would then take the average over the fitted values of the polynomial, and then we require those averages to stay low.

- The time between the inflection point and the top varies from peak to peak. As as result, the maximal possible fit range (time between the inflection point and the top minus 5 minutes we need for the ahead predictions) also varies. For better estimations of the location of the top, it would be better if we used this maximal possible fit range dependent on the peak, rather than a fixed fit range (the minimum of the maximal possible fit ranges of the peaks so far). We notice that this maximal possible fit range cannot be determined by just subtracting 5 minutes from the time between the inflection point and the actual top: during real-time predictions, we do not know the actual top beforehand. Instead, we need a method that can determine the maximal possible fit range without using the actual top. One possible option relies on the observation that the tide approximately follows a sinusoid. For a perfect sinusoid, the time difference between the maximal and the inflection point equals the time difference between the minimum and that same inflection point. Thus we might be able to estimate the maximum possible fit range by the time difference between the minimum and that inflection point minus 5 minute prediction time.

- In this research paper, fitting of the second model (with the Erlang distribution) is splitted into two parts: first, the parameters of the trend are estimated by fitting solely the trend (so without the short-term oscillations). Secondly, the full model is fitted including the short-term oscillations. This splitting is done because finding the optimal values for all parameters (so the trend as well as the short-term oscillations) at once takes too much time for the algorithm. However, in the research paper also a method of omitting the phases as non-linear parameter was discussed. One possible future research would rely on investigating whether it is possible to use this reduction to fit the full model at once. It would be interesting to know if this approach leads to better fits and predictions than the original approach.

- In Section 6.3 we also discussed one method reduce the non linear parameter $\lambda$ in the second model, but for that method we needed to know the location of the top. For real-time predictions we do not know this location, however. Another method that could be tried is to once fit without reduction of $\lambda$, and use the found estimate of the location of the top (see Section 7.3) to do use the reduction of the $\lambda$ from the

second fit on wards. This way time gain is obtained from that moment on. This might however be at the expense of the quality of the fits.

- Yet another method to lower the fit time of the second model might be to increase the time steps in the ranges of starting values of the non-linear parameters, so that there are less total non-linear squares optimizations the software has to solve. This might however be at the expense of the quality of the fits.
- In our current models we over and over used three sinusoids to model the short-term oscillations, as this is the suggested amount by the Dutch Ministry of Infrastructure and Water Management. A way to further reduce the total error $\varepsilon$ however, might be to include more than three sinusoids to model the short-term oscillations, to better model the short term behaviour.
- In our current statistical models we only used one exploratory or independent variable, namely the time. To reduce the total error it might be a good idea to include other exploratory variables as well, most notably the wind. This might be a good idea as the wind is one of the main sources of high(er) sea waves.

# Bibliography

Rex Bolsius. Modelling the sea level near the eastern scheldt storm surge barrier. Master's thesis, Eindhoven University Of Technology, May 2018.

Bart Dassen. Short-term water predictions for the dutch storm surge barriers. Master's thesis, Eindhoven University Of Technology, July 2016.

M. Ramaekers and J. Michels. Data assignment sea water. Report, Eindhoven University Of Technology, January 2019.

Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Its Applications*. Springer Science and Business Media, fourth edition, 2016.

Wikipedia. Oosterscheldekering. `https://nl.wikipedia.org/wiki/Oosterscheldekering`, 2019. Accessed on 07-10-2019.