STATISTICAL ANALYSIS OF OCEANOGRAPHIC DATA: A COMPARISON BETWEEN STATIONARY AND MOBILE SEA LEVEL GAUGES

MARCUS POSADA

Master's thesis
2014:E12

LUND UNIVERSITY

Faculty of Engineering
Centre for Mathematical Sciences
Mathematical Statistics

CENTRUM SCIENTIARUM MATHEMATICARUM

TYP AV DOKUMENT		DOKUMENTBETECKNING
☒ Examensarbete	☐ Kompendium	
☐ Delrapport	☐ Rapport	LUTFMS—3242—2014

INSTITUTION

Matematikcentrum. Matematisk statistik, Lunds universitet, Box 118, 221 00 LUND

FÖRFATTARE

Marcus Posada

DOKUMENTTITEL OCH UNDERTITEL

Statistical Analysis of Oceanographic Data: A Comparison between Stationary and Mobile Sea Level Gauges

SAMMANFATTNING

This Master's thesis project concerns developing statistical methods to examine sea level data provided by the Swedish Meteorological and Hydrological Institute (SMHI). The data comes from sea level gauges located in the harbours of Uddevalla, Ängelholm and Åhus along the Swedish coast. These three gauges are mobile, as compared to the permanent sea level gauges used by the SMHI at various points along the coast. The mobile gauges were placed during 2010 and the goal of this project is to analyse the data they have provided. This is done in several steps as outlined below.

Initially, a comparison is performed between the extreme sea levels at the three mobile gauge locations and permanent gauges placed nearby. This analysis results in estimates of five-year return levels for the differences between the mobile and nearby permanent gauges. It turns out that the difference is roughly 50 cm. The second analysis involves studying the relationship between the sea levels at the paired stations using linear regression. As a final analysis these models are diagnosed and extended to include parameters for wind speed, wind direction and atmospheric pressure.

The largest single objection to the validity of the regression models is arguably the natural temporal dependence in the data which indicates that time series models could be more suitable. Such models are fitted to the dataset where fractional integration is applied to handle long-term memory and GARCH models are implemented for conditional heteroscedasticity. In addition, harmonic tidal analysis is also performed.

NYCKELORD

DOKUMENTTITEL OCH UNDERTITEL - SVENSK ÖVERSÄTTNING AV UTLÄNDSK ORIGINALTITEL

Statistisk analys av oceanografiska data: en jämförelse mellan stationära och mobila havsnivåmätare

UTGIVNINGSDATUM	ANTAL SID	SPRÅK
år 2014 mån 3		☐ svenska ☒ engelska ☐ annat

ÖVRIGA BIBLIOGRAFISKA UPPGIFTER	ISSN
	ISBN
	2014: E12

Signature Marcus Posada Date 24/3 -14

Abstract

This Master's thesis project concerns developing statistical methods to examine sea level data provided by the Swedish Meteorological and Hydrological Institute (SMHI). The data comes from sea level gauges located in the harbours of Uddevalla, Ängelholm and Åhus along the Swedish coast. These three gauges are mobile, as compared to the permanent sea level gauges used by the SMHI at various points along the coast. The mobile gauges were placed during 2010 and the goal of this project is to analyse the data they have provided. This is done in several steps as outlined below.

Initially, a comparison is performed between the extreme sea levels at the three mobile gauge locations and permanent gauges placed nearby. This analysis results in estimates of five-year return levels for the differences between the mobile and nearby permanent gauges. It turns out that the difference is roughly 50 cm. The second analysis involves studying the relationship between the sea levels at the paired stations using linear regression. As a final analysis these models are diagnosed and extended to include parameters for wind speed, wind direction and atmospheric pressure.

The largest single objection to the validity of the regression models is arguably the natural temporal dependence in the data which indicates that time series models could be more suitable. Such models are fitted to the dataset where fractional integration is applied to handle long-term memory and GARCH models are implemented for conditional heteroscedasticity. In addition, harmonic tidal analysis is also performed.

Preface

This Master's thesis project was done by me, Marcus Posada at the Centre for Mathematical Sciences, Mathematical Statistics, at the Faculty of Engineering (LTH), Lund University, in collaboration with the Oceanographic Warning & Forecasting Service at the Swedish Meteorological and Hydrological Institute, SMHI. The questions at hand were formulated by the Oceanographic Warning & Forecasting Service at SMHI, but the choice of appropriate statistical methods was left up to me and my supervisor, Nader Tajvidi at LTH.

I would like to thank my supervisor, Nader Tajvidi, for much valuable support and help during the project. I would also like to thank Robert Olsson at SMHI for proposing the project to me. At SMHI I would also like to thank Patrik Strömberg who has been my main contact and has given much valuable input as well as Fredrik Waldh and Hans Bengtsson, for answering questions and providing some of the data used during the project. I also thank Christina Nilsson-Posada for providing input regarding the language in the report. Finally, I thank my fiancée Heléne Alpfjord, who has given me invaluable support during the entire project and without whose help neither this report, nor this project as a whole would be as good.

Contents

1 Introduction

Through experience, the on duty oceanographers (Vakthavande Oceanografer) at the Swedish Meteorological and Hydrological Institute (SMHI) knows that at certain points along the Swedish coast the models used for forecasting sea levels sometimes predict less extreme values than are later observed. These points of interest tend to lie in bays, where factors such as water depth, wind speed and wind direction can make for significantly different sea level behaviour, compared to nearby points along the coast but outside the bays (Strömberg 2012). This pertains to both high and low sea water levels.

SMHI has 23 permanent sea level gauges placed along the coast, from Kalix, close to the Finnish border, to Kungsvik, right along the border to Norway. The longest still active gauging position is Stockholm, where sea level data have been recorded from 1889 onwards (smhi.se 2013f). In addition to these gauges, SMHI also has three mobile sea level gauges placed at Uddevalla, Ängelholm and Åhus, all positions that SMHI consider to be oceanographically interesting areas. These mobile gauges were placed during 2010 and are still operational.

Now, SMHI wish to relate the sea level data from the mobile gauges with the sea level data from nearby permanent stations in order to eventually be able to increase the reliability of the oceanographic forecasting and warning service. The permanent stations that have been chosen for comparison are for Uddevalla: Smögen, for Ängelholm: Viken and for Åhus: Simrishamn and also (possibly) Kungsholmsfort. They also wish to study the effects of wind speed, wind direction and atmospheric pressure on the sea levels.

The report is structured into six sections besides the introduction, as well as an appendix, which is divided into three sections. Section 2 details the geographical locations of the gauges in question, provides a brief overview of some factors that affect sea levels along the Swedish coast, and defines the warning levels used by the Oceanographic Warning & Forecasting Service at SMHI. In Section 3, an initial analysis is done on the data provided by SMHI. It details information regarding missing data, some statistical characteristics, and correlations between the data sets. Section 4 is theoretical and in it, an outline of extreme value theory is given, with a focus on the GEV distribution, as well as regression analysis, including model design and diagnostics. Finally, the theory section also outlines time series analysis models, focusing on variations of the ARMA model, including fractional integration and autoregressive conditional heteroscedastic models. A brief account of tidal harmonic analysis is also given.

In Section 5, the results of applying the theory and methods described in

Section 4 to the sea level data sets are presented. The extreme value theory is applied to the differences between the paired stations. This is done with the goal of estimating how large the differences could get, with confidence bounds. The regression analysis is used to study the relationships between the data sets for the mobile gauges and their permanent counterparts. These regression models are diagnosed and then extended to include the wind speed, wind direction and atmospheric pressure data. The largest objection to the validity of a regression model is the natural temporal dependence in the data. Thus, time series models are built for the data. These are initially quite large and contain structures to handle long term memory and conditional heteroscedasticity. Harmonic tidal analysis is performed, in order to reduce the size of the models.

The Results section is followed by Section 6, which gives details regarding the programming languages and packages used in the project. Section 7 is Summary and Discussion, in which the results are summarised and possible future steps concerning the data from the mobile gauges are suggested.

2 Background

2.1 Locations of the mobile gauges

The mobile gauges that are the focus of this project are, as stated in the Introduction, located in Uddevalla (in decimal degrees c. N 58°.34' E 11°.89'), Ängelholm (c. 56.27 12.82) and Åhus (c. 55.93 14.33). They are marked in Figure 1.

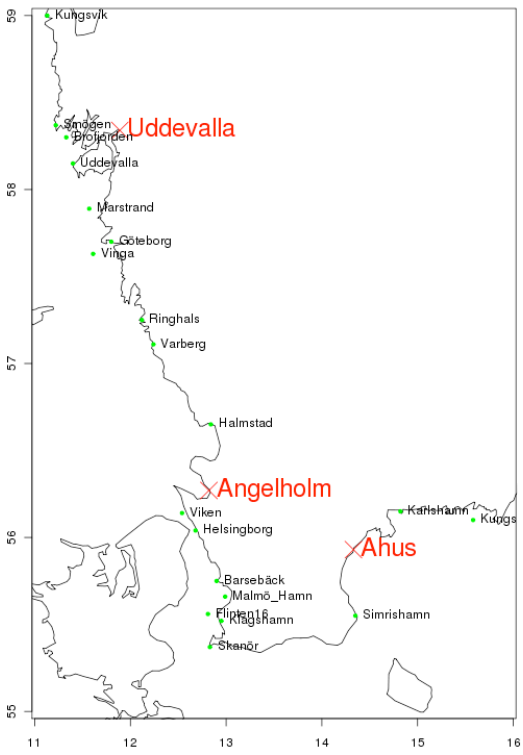

Figure 1: The locations of the three mobile gauges are marked with red. The green dots represent permanent sea level gauges as well as a few points that only represent model values. Image from Strömberg 2012, p. 4.

For technical specifications regarding the gauges, see Strömberg 2012.

2.2 Locations of the permanent gauges

The three mobile gauging stations have each been paired with a permanent station, as described in the Introduction. These permanent stations are all

located close to their respective mobile gauges. In Figure 1 it can be seen that Smögen, Viken and Simrishamn lie close to Uddevalla, Ängelholm and Åhus, respectively.

2.3 Reference frames

Sea water level data from SMHI are given in the reference frame RH2000 (Rikets höjdsystem 2000) which is the national Swedish vertical reference frame since 2005. In order to make the different gauging stations comparable, the sea level data needs to be corrected for relative sea level changes, i.e. for isostasy and eustasy (Swe., "landhöjning" and "havsnivåhöjning") (smhi.se 2013c).

The annual Mean Sea Level (MSL) is a value that is calculated through a regression of many annual average sea levels. To be able to calculate the regression line to a good enough precision, at least 30 years of data is needed (Strömberg 2012; smhi.se 2013a).

The aforementioned correction is done by adding a constant that represents the MSL for the current year and gauging station to the RH2000 data. Since the time series for the mobile stations are too short to accurately determine the MSL, the corrections for the mobile stations are done by using the MSL for the nearest gauging station that has at least 30 years of observations. For the equations that describe the MSL, see Appendix A.

All sea level data in this project, from both mobile and permanent stations, have been corrected for mean sea level changes according to Appendix A.

2.4 Factors that affect sea levels along the Swedish coast

Of course, several factors influence the sea water levels along the Swedish coast. Examples are atmospheric pressure, wind direction, wind speed, lunar and solar tides, water density, depth of the basin, isostasy and eustasy. The last two have been corrected for in the data used in this project, as described in Section 2.3. Along the Swedish coast the most influential factors are the wind over the North Sea and the Baltic Sea as well as atmospheric pressure conditions (smhi.se 2013b).

The geography of the coast can naturally influence the sea levels a lot. One example of this is the gauge station in Uddevalla harbour, seen in Figure 2. It is not far-fetched to assume that the water levels in Uddevalla harbour can differ greatly from the water levels measured in Smögen, located further out, due to for example water stowage in Havstensfjorden during certain wind/atmospheric pressure conditions.

Figure 2: There is a mobile gauge station in Uddevalla harbour, near the upper right corner of the map. The location of Smögen, where the corresponding permanent station is, is marked by the red pin on the map. Image: © OpenStreetMap contributors (openstreetmap.org).

The sea levels along the Swedish coast often exhibit clear seasonality as the winds and atmospheric pressure changes with the seasons. The sea levels can vary very much between different points along the coast over short time periods, but at but half-year scale averages are roughly equal at all of SMHI's gauging stations.

Due to predominantly southwesterly winds during autumn, which press water towards the west coast, sea levels are typically high. During winter the sea levels exhibit stronger variation due to stronger winds. The most

common case during winter is that the season is dominated by low atmospheric pressure and southwesterly winds, causing high sea levels. During spring and summer the sea levels typically drop and have smaller variations, due to higher pressure and weaker winds. Southwesterly winds can cause high sea water levels along the west coast and the northern Baltic Sea coast while low water levels are recorded in Öresund (smhi.se 2013d).

Higher atmospheric pressure generally leads to lower water levels. According to calculations done by SMHI, an increase of the atmospheric pressure by 1 hPa gives a 1 cm drop in sea water levels. Since the atmospheric pressure in Sweden typically lies between 950 and 1050 hPa, pressure causes sea water level variations between +63 and −37 cm, though this effect is hard to observe in practice (smhi.se 2013e).

Finally, the sea levels are also influenced by deterministic tides. In the Baltic, the tide is barely noticeable with an effect of only a few centimetres. This is due to its small size and narrow entrance at Öresund. In Kattegatt and Skagerack, on the other hand, the amplitude of the tides can reach 40 cm, under the right conditions. The tide along the Swedish coast is semidiurnal with a dominating period of 12 hours and 25 minutes, which is caused by the moon's gravitational pull on Earth. Of course, the tide is also affected by other factors such as the gravitational pull of the sun, the earths rotation, variations in the slope of their orbits and the local geography (smhi.se 2013g).

In analysis of oceanographical data, it is common to separate the tidal signal from the non-deterministic noise signal. The tidal signal can then either be studied by itself or discarded(Pawlowicz, Beardsley, and Lentz 2002).

2.5 The Oceanographic Warning & Forecasting Service

One of the responsibilities of the Oceanographic Warning & Forecasting Service at SMHI is to issue and convey warnings when sea levels along the Swedish coast are predicted to be especially high or low. These warnings are divided into two classes, Class 1 and Class 2, with Class 2 being more severe. The exact warning levels for the locations of interest are given in Table 1.

Generally, warnings of Class 1 are considered to have much informative value, pose some risk to the public and may disrupt some public services. Class 2 warnings indicate an oceanographic development that could pose a danger to the public, as well as major property damage and major disruptions of important public services. Warnings for high sea levels are of interest since buildings, roads and quays risk being flooded. Also, installations such as water treatment plants can experience problems. Low sea levels are mainly

	Class 1		Class 2	
	Low	High	Low	High
Uddevalla		80	−100	120
Ängelholm	−60	80	−100	120
Åhus		80	−100	120

Table 1: Warning levels for the three locations of interest, in cm. They are the same for all of the locations, except the Class 1 warning for low sea level in Ängelholm, which is −60, but does not exist for the other two locations.

of interest to the maritime industry, as they can result in groundings and ships having problems entering certain harbours (Jönsson and Olsson 2013).

A warning should cover an area of 1000 m^2 and reach within 10 cm of the warning level (as defined in Table 1) to be regarded as correct.

Throughout this report, whenever warning levels are mentioned, these will be the levels in question unless stated otherwise.

3 Data

3.1 Initial data analysis

3.1.1 Quality of data

In contrast to the data from the permanent gauges, the data from the mobile gauges has not been quality controlled by SMHI. Thus, there are instances of both missing data and clear measurement error.

In the Ängelholm data series, one obviously incorrect measurement has been found, on the 15th of March 2013. This is shown in Figure 3. In the Ängelholm data there are 1787 hourly values missing (including the one obviously incorrect measurement found). Of these, 63 % come from a period of 46 days in April and May 2012, when the gauge was apparently not working.

In the Uddevalla data there are 945 hourly measurements missing, of which 92% come from a period of 36 days during the summer of 2012 when the gauge was not working. Also, one obviously incorrect measurement has been found. It is also marked as missing.

In the Åhus data, there are 29 hourly measurements missing and in the data series from the permanent gauging stations, no data is missing.

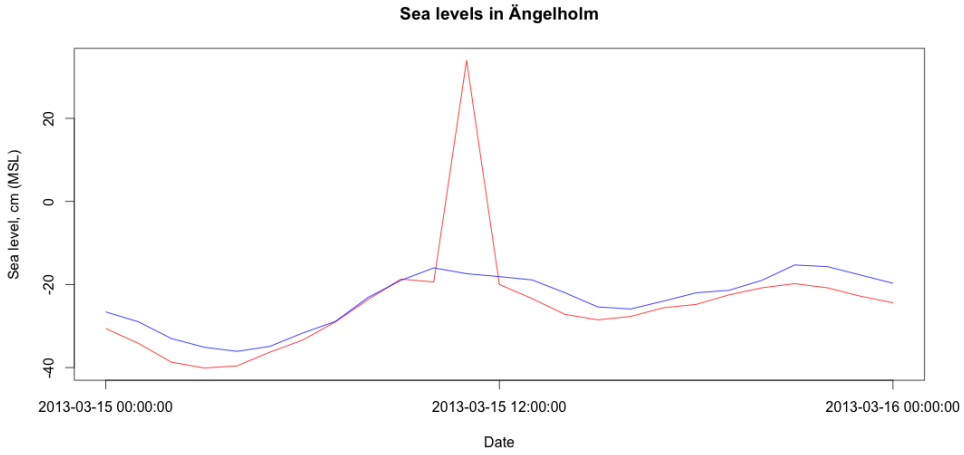

Figure 3: Sea levels in Ängelholm (red line) plotted together with sea levels in Viken (blue line) on the 15th of March 2013.

3.1.2 Characteristics of the data

The minimum, mean, maximum and standard deviation of each of the time series studied in this project are shown in Table 2. Looking at the three pairs of stations separately; one can see that the range of the datasets is larger for the mobile stations than for their paired permanent stations, in all three cases. Also, the standard deviations are similar in all cases, though it is somewhat higher in Uddevalla than in the other locations.

	Minimum:	Mean:	Maximum:	Standard deviation:
Uddevalla (mob.)	-80.15	2.01	150.65	26.91
Smögen	-65.1	4.32	122.9	22.82
Ängelholm (mob.)	-108.1	-4.35	178	23
Viken	-98.3	-1.15	156.1	20.80
Åhus (mob.)	-125.3	-9.76	88.72	21.26
Simrishamn	-113.9	-13.18	89.62	20.72

Table 2: Mean, maximum, minimum and standard deviation for the three datasets from the mobile gauges (Uddevalla, Ängelholm and Åhus) as well as from the three permanent ones (Smögen, Viken and Simrishamn). All values in the table are in cm.

All three of the time series from the mobile gauges are shown in Figure 4, together with the nearby permanent stations. In all three plots, the mobile data is shown in red, while the data from the nearby permanent gauges is plotted in blue. Looking closer at the two topmost plots, one can easily see the two longer periods where data is missing; in Uddevalla from the summer of 2012, in Ängelholm from the spring of 2012. All the plotted datasets are clearly heteroscedastic (i.e. the data lacks homogeneity of variance).

The Åhus and Simrishamn time series exhibit quite different general behaviours compared to the other two locations. This difference is assumed to be due to the lack of pronounced tides in the Baltic Sea, as discussed in Section 2.4.

The yellow and orange lines in Figure 4 show where the Class 1 and Class 2 warning levels are, for both high and low water levels. The warning levels are given in Section 2.5.

Another way of visualising the influence of the tide along the west coast is to look at the autocorrelation plots in Figure 5. The maximum lag in the plots is a week.

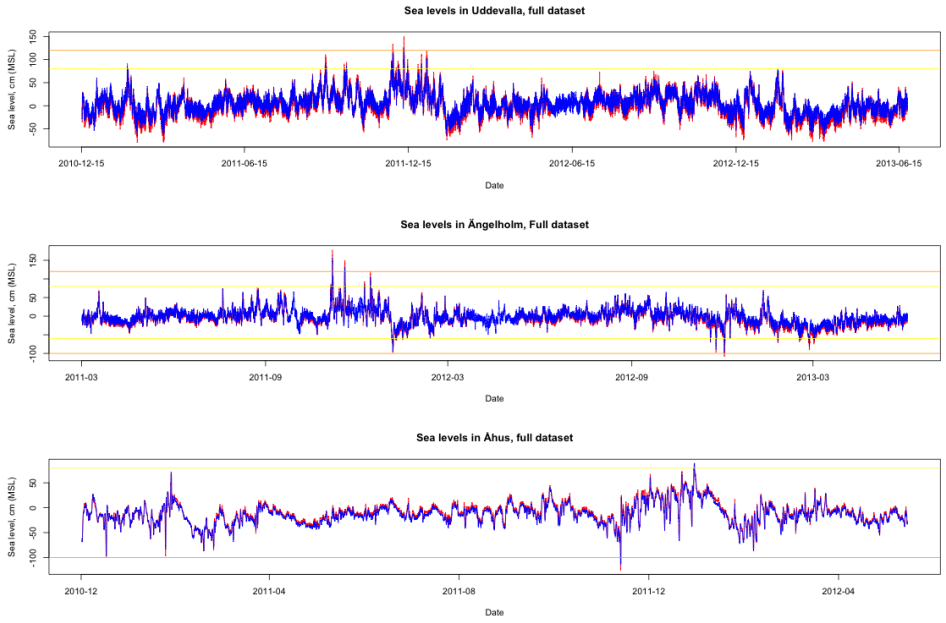

Figure 4: Top: data from Uddevalla in red, ovarlayed with data from Smögen in blue. Middle: data from Ängelholm in red with data from Viken in blue. Bottom: data from Åhus in red with data from Simrishamn in blue. The yellow and orange lines show Class 1 and Class 2 warning levels respectively, for both high and low water levels, where applicable.

Looking at the histograms and qq-plots presented in Figure 6, it can be seen that none of the three data sets from the mobile stations appear to come from a normal distribution. The two topmost plots (Uddevalla) show that the data is slightly skewed to the right. The four lower plots, corresponding to Ängelholm and Åhus, show that the data has longer tails than would be expected if it had been normally distributed, possibly suggesting a Student's t-distribution.

3.1.3 Correlations

There are strong correlations between the data from the mobile gauges and their paired permanent gauges. For Uddevalla-Smögen the (Pearson's) correlation is $\rho_{\mathrm{U,S}} = 0.971$, for Ängelholm-Viken it is $\rho_{\mathrm{Ä,V}} = 0.982$ and for Åhus-Simrishamn it is $\rho_{\mathrm{Å,S}} = 0.991$. Since the relationships are all linear, as can be seen in Figure 7, Pearson's correlation coefficient describes the relationship between the data sets well and more robust dependence measures,

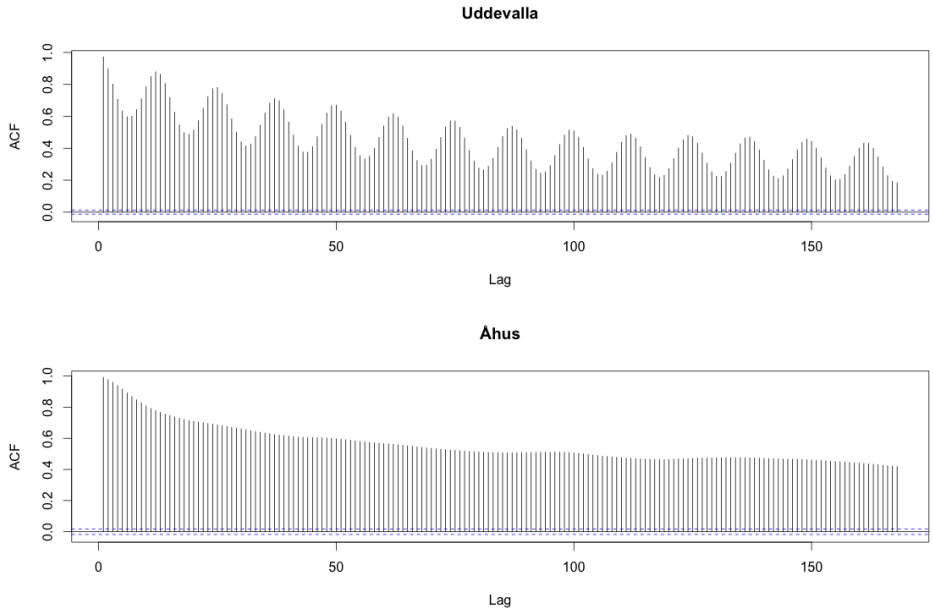

Figure 5: Autocorrelation plots for Uddevalla and Åhus. The semidiurnal influence of the tide is clearly visible in the topmost plot. The lags are hours and the maximum lag is one week.

such as Kendall's τ or Spearman's ρ, are not needed.

Looking closer at at, for example, the Uddevall-Smögen plot, it can be seen that there are a few instances where the water level in Uddevalla lies above the Class 2 warning level, even though the water levels in nearby Smögen have not exceeded the warning level.

3.2 Wind and atmospheric pressure data

Also available is data concerning the wind speed, wind direction and air pressure at Uddevalla, Ängelholm and Åhus. Figure 8 shows a wind rose plot for Uddevalla. It is clear that a large proportion of the wind in Uddevalla comes from southwest. The wind roses for Ängelholm and Åhus show predominantly southerly and westerly winds, respectively. The atmospheric pressure data is also hourly and is given in hPa.

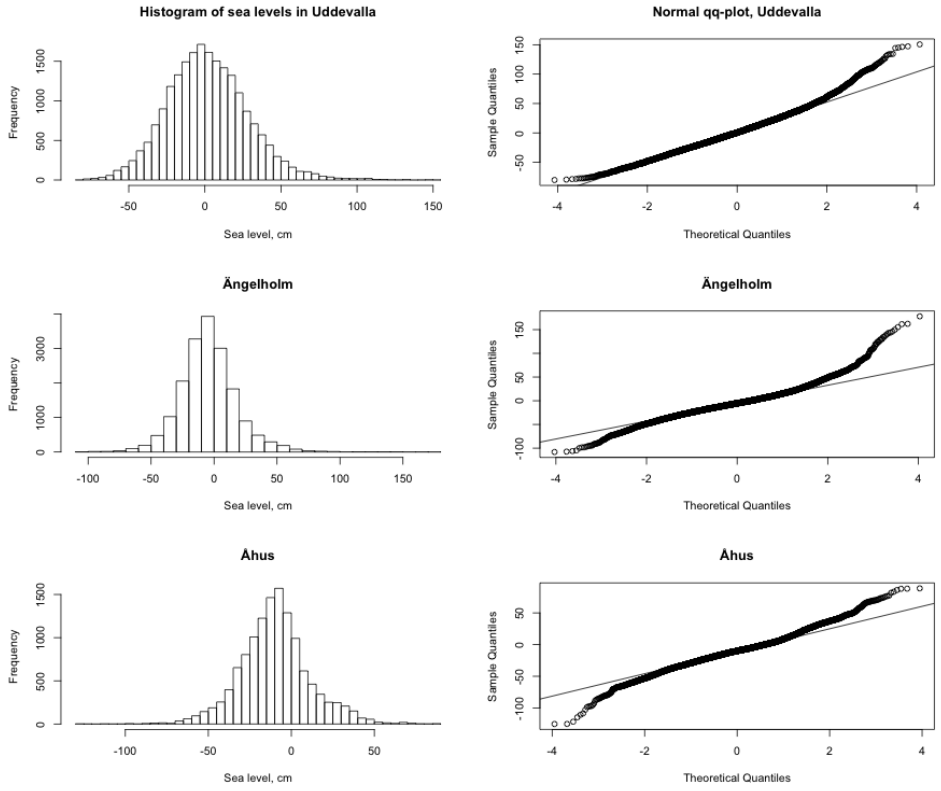

Figure 6: The rows correspond to Uddevalla, Ängelholm and Åhus. The two columns show histograms and normal probability plots, respectively.

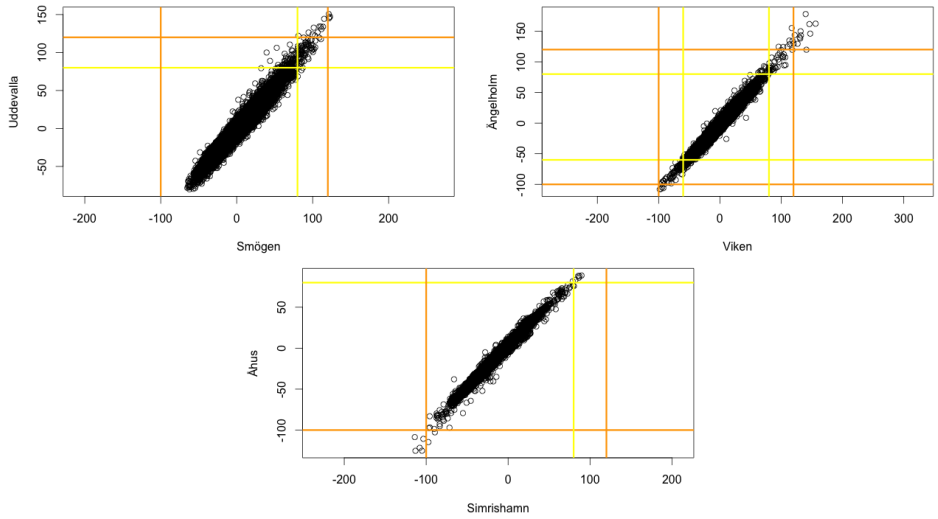

Figure 7: Scatterplots for the three pairs of datasets studied. The data from the mobile gauges are on the y-axes while the data from the permanent gauges are on the x-axes. The yellow and orange lines show Class 1 and Class 2 warning levels respectively, for both high and low water levels, where applicable.

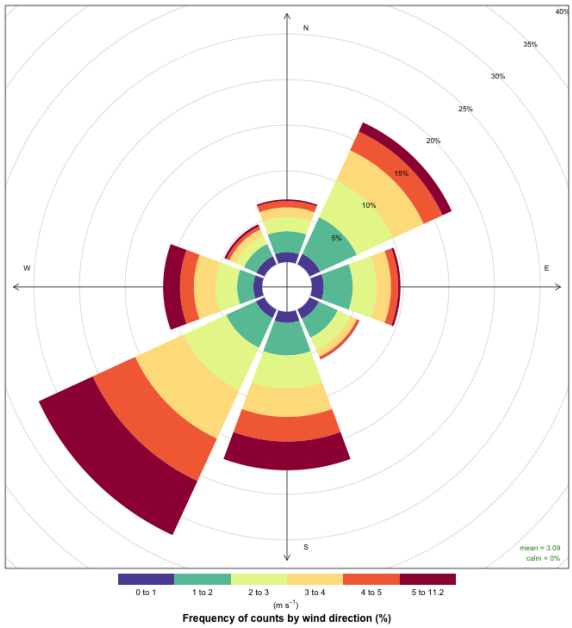

Figure 8: A wind rose plot for Uddevalla. The plot shows the frequency of winds blowing from eight different directions (clock-wise: N, NE, E, SE, ..., NW). The length of each spoke shows the frequency of each of these directions. The colour bands inside each spoke show wind speed ranges, in m/s. The longest spoke shows that roughly 27 % of the wind in Uddevalla comes from the southwest.

4 Theory

This Section is divided into three main parts: extreme value analysis in 4.1, regression analysis in 4.2 and time series analysis in 4.3. The extreme value part covers classical extreme value theory and defines return levels and return periods. The regression analysis part outlines simple linear regression and also covers multiple linear regression and regression diagnostics. Finally, the time series analysis section describes the autoregressive (AR) and the moving average (MA) models, as well as their combination, the ARMA model. The section also contains a method for dealing with non-constant variance, the generalised autoregressive conditional heteroskedasticity (GARCH) model, and long term memory, the autoregressive fractionally integrated moving average (ARFIMA) model.

4.1 Extreme value analysis

The main objective of extreme value analysis is to quantify the behaviour of a process at *extreme* levels. What is considered extreme has to be decided, subjectively, for every case studied. Extreme value theory is a framework that makes it possible to make extrapolations about the characteristics of a process beyond the most extreme observations that have been made.

4.1.1 Generalised Extreme Value (GEV) distribution

The classical extreme value theory focuses on the distribution of block maxima. Block maxima are defined as $M_n = \max(X_1, \ldots, X_n)$. Here, $X_1, \ldots, X_n$ is a sequence of independent and identically distributed (iid) random variables. E.g. if the X_i's are hourly measurements, M_{168} would be the weekly maximum. Since the population distribution is usually not known in practical applications, but needs to be estimated in order to describe the behaviour of the block maxima.

The extremal types theorem (also known as the Fisher-Tippett-Gnedenko theorem, the extreme value theorem or the convergence of types theorem) was first formalised by Gnedenko in 1948. It states that if M_n can be rescaled with a sequence of constants $\{a_n > 0\}$ and $\{b_n\}$ so that its distribution does not converge to a point mass, the distribution must belong to one of only three possible families. In the words of Coles (2001),

Theorem 4.1.1. *If there exist sequences of constants* $\{a_n > 0\}$ *and* $\{b_n\}$ *such that*

$$\mathrm{P}\{(M_n - b_n)/a_n \leq z\} \to G(z) \text{ as } n \to \infty,$$

where G is a non-degenerate distribution function, then G belongs to one of the following families:

$$\begin{aligned}
\text{I} : G(z) &= \exp\left\{-\exp\left[-\left(\frac{z-b}{a}\right)\right]\right\}, \ -\infty < z < \infty \\
\text{II} : G(z) &= \begin{cases} 0, & z \leq b, \\ \exp\left\{-\left(\frac{z-b}{a}\right)^{-\alpha}\right\}, & z > b, \end{cases} \\
\text{III} : G(z) &= \begin{cases} \exp\left\{-\left[-\left(\frac{z-b}{a}\right)^{\alpha}\right]\right\}, & z < b, \\ 1, & z \geq b, \end{cases}
\end{aligned}$$

for parameters $a > 0$, b and, in the case of families II *and* III*, $\alpha > 0$.*

The three families are called Gumbel, Fréchet and Weibull, respectively. Unfortunately, the extremal types theorem leaves the possibly very difficult choice of which distribution family to choose for the problem at hand. To get around this choice, von Mises in 1954 and Jenkins in 1955 independently designed a new family, the Generalised Extreme Value (GEV) distribution,

$$G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\sigma}\right\}.$$

The GEV distribution has three parameters, a location parameter, $-\infty < \mu < \infty$, a scale parameter $\sigma > 0$, and a shape parameter, $-\infty < \xi < \infty$. It is defined on $\{z : 1+\xi(z-\mu)/\sigma > 0\}$. This distribution has all three families described in Theorem 4.1.1 as special cases. The GEV distribution is equal to the three different distribution families for values of the shape parameter that are, $\xi = 0$, $\xi > 0$ and $\xi < 0$, respectively. For a sketched proof of the extremal types theorem, see Coles (2001) and for a more formal justification, see Leadbetter, Lindgren, and Rootzén (1983).

The apparent need of knowing the normalizing constants b_n and a_n makes estimating the parameters of G a difficult problem. But this problem can be avoided in practice, by estimating the parameters of an equivalent distribution. Since

$$\mathrm{P}\{(M_n - b_n)/a_n \leq z\} \approx G(z),$$

is equivalent to

$$\mathrm{P}\{M_n \leq z\} \approx G\{(z - b_n)/a_n\} = G^*(z).$$

Thus, the parameters of G^* can be estimated instead of the parameters of G and the distribution of M_n will be approximated (Coles 2001).

As previously stated, M_n is a sequence of block-maxima, where each block is an observational sequence of length n. The typical block-size to use is one year, chosen for the simplicity of interpretating the results and because data is very often available annualy, but the theory works equally well for other block-sizes.

4.1.2 Return levels and return periods for GEV distributions

Return levels, z_p are quantiles of the extreme value distribution, i.e. levels at which $G(z_p) = 1 - p$. The return level z_p is a level such that it will be exceeded on average once every $1/p$ periods (here, the period can be e.g. one year, six months or one week). $r_p = 1/p$ is called the return period (Coles 2001).

If, for example, the return level of interest is the five year return level and the data involved are divided into two-week blocks, the return period will be $r_p = 5 \cdot 52/2 = 130$ and p will be 0.0077. Hence, z_p will be smaller than the two-week maximum with a probability of 0.0077 and z_p will be exceed on average once every 130 periods, i.e. once every $130/26 = 5$ years.

Expressions that estimate z_p can be acquired from the GEV definition through algebra. They are

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}[1 - \{-\log(1-p)\}^{-\xi}], & \xi \neq 0, \\ \mu - \sigma \log\{-\log(1-p)\}, & \xi = 0. \end{cases}$$

In order to create confidence intervals for the return levels, the variance of $\hat{z}_p$ needs to be estimated. As described in e.g. Coles (2001), the delta method estimate of this variance is

$$Var(\hat{z}_p) \approx \nabla z_p^T V \nabla z_p.$$

Here, V is the variance-covariance matrix of the estimated GEV parameters and

$$\begin{aligned} \nabla z_p^T &= \left[\frac{\partial z_p}{\partial \mu}, \ \frac{\partial z_p}{\partial \sigma}, \ \frac{\partial z_p}{\partial \xi} \right] \\ &= \left[1, \ -\xi^{-1}(1 - y_p^{-\xi}), \ \sigma\xi^{-2}(1 - y_p^{-\xi}) - \sigma\xi^{-1} y_p^{-\xi} \log(y_p) \right], \end{aligned}$$

evaluated using the estimated GEV parameters.

4.2 Regression analysis

The goal of linear modelling is to describe how the mean of a dependent variable, denoted by Y, changes with varying conditions. These conditions

are described by independent variables, also called predictor or explanatory variables, denoted by X_i.

This Section of the report is divided into five subsection, each either describing an extension to the simple linear model, as it will be described in Subsection 4.2.1, or a methodology to diagnose the fitted models.

4.2.1 Simple linear regression

The simplest linear regression model only has one independent variable, X. The linear regression describes how the mean of the dependent variable changes linearly with the independent variable,

$$E(Y_i) = \beta_0 + \beta_1 X_i.$$

To compensate for the fact that the observation that we have available deviates from the population mean, random errors ϵ_i, are added. This gives the classic simple linear model,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i. \tag{1}$$

Some assumptions have been made at this stage: the observed values of X are assumed to be known and the observations of the dependent variable are assumed to be random observations from populations of random variables with means $E(Y_i)$. Also, the errors, ϵ_i, are identically and independently distributed (i.i.d.) with variance σ^2. Thus, also the Y_i's are pairwise independent and have a common constant variance, σ^2.

In order to estimate the values of the two parameters β_0 and β_1 using the two data sets Y and X a method is needed. The most common method is called least squares estimation. The least squares solution gives the smallest possible sum of squared deviations between the observations and the estimated line. I.e., least squares estimation finds the numerical estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimise the sum of squares of the residuals:

$$SS(Residual) = \sum_{i=1}^{n} (Y_i - \hat{Y}_1)^2,$$

where $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$. The expressions that describe $\hat{\beta}_0$ and $\hat{\beta}_1$ can be acquired using calculus or looking at a statistics or regression textbook, such as Rawlings, Pantula, and Dickey (2001) or Blom et al. (2005). Each point on the regression line, $\hat{Y}_i$, has two interpretations. They can be seen as either predictions of the values of the dependent value that can be obtained for a

future observation of the independent variable or as least squares estimates of the population mean for a specific value of X.

4.2.2 Multiple linear regression

The simple linear regression model, as stated in section 4.2.1, can be extended to include more than one independent variable:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i.$$

Now, instead of two parameters, $p' = p + 1$ parameters need to be estimated using p independent data sets. Using the same assumptions as in the case with only one dependent variable, the $SS(Res)$ can be minimised, giving p' $\hat{\beta}_j$ estimates.

This notation quickly becomes tedious with many parameters. The linear model can written more conveniently using a matrix notation,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{2}$$

Here, $\mathbf{X}$ is an $(n \times p')$ matrix:

$$\begin{bmatrix} 1 & X_{11} & X_{12} & X_{13} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & X_{23} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & X_{n3} & \cdots & X_{np} \end{bmatrix}$$

while $\mathbf{Y}$, $\boldsymbol{\epsilon}$ and $\boldsymbol{\beta}$ are $(n \times 1)$, $(n \times 1)$ and $(p' \times 1)$ vectors, respectively. As in the one independent variable case, the $\mathbf{X}$ matrix contains values that are assumed to be known constants and $\boldsymbol{\epsilon}$ is a random vector of independent random variables where each $\epsilon_i \in N(0, \sigma^2)$. Thus, the elements of $\mathbf{Y}$ are assumed to be independent and normally distributed and $\mathbf{Y} \in N(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}\sigma^2)$. Similarly to the one independent variable case, the fitted values are $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ and the model residuals are $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$.

4.2.3 Analysis of variance

The total, uncorrected, sum of squares of the observed values is defined as $SS(\text{Total}_{\text{uncorr}}) = \sum_{I=1}^{n} Y_i^2 = \ldots = \sum_{I=1}^{n} \hat{Y_i}^2 + \sum_{I=1}^{n} e_i^2$. The first sum is called $SS(\text{Model})$ and the second $SS(\text{Residual})$. As their names indicate, they describe a division of the total uncorrected sum of squares into two parts, one that accounts for the model ($\mathbf{X}\boldsymbol{\beta}$) and one that accounts for the residual that the model fails to describe. $SS(\text{Model})$ can be further divided,

$SS(\text{Model}) = n\bar{Y} + \hat{\beta}_1^2 \sum_{i=1} n(X_{1i} - \bar{X}_1)^2 + \ldots$. The first term, $n\bar{Y}$, is the sum of squares correcting for the mean, $SS(\mu)$. Removing it from $SS(\text{Total}_{\text{uncorr}})$ gives the expression for the total, corrected, sum of squares: $SS(\text{Total}_{\text{corr}}) = SS(\text{Regr}) + SS(\text{Residual})$. $SS(\mu)$ is the sum of squares of a model only containing a constant term β_0, without any independent variables. Since the question at hand is usually how much the added regression terms can explain the variation of Y compared to a model that is only a mean, $SS(\text{Regr})$ is of interest. It expresses the information given by including the independent variables in the model.

4.2.4 Variable selection

Once the parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ have been found, the next question is: is the linear dependence significant, i.e. is $\hat{\beta}_1$ significantly different from zero? The typical method for answering this question is the t-test. It test the null hypothesis, H_0, that $\beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. If H_0 is true,

$$\frac{\hat{\beta}_1 - 0}{\sigma / \sqrt{\sum_{1=1}^{n} (X_i - \bar{X})^2}} \in N(0, 1)$$

and since the true value of σ is unknown in practice,

$$t = \frac{\hat{\beta}_1 - 0}{s / \sqrt{\sum_{1=1}^{n} (X_i - \bar{X})^2}} \in t(n-2).$$

The test rejects the null hypothesis at significance level α if $|t| > |t_{\alpha/2}(n-2)|$ (Rawlings, Pantula, and Dickey 2001). Here, s^2 is an estimate of σ^2:

$$s^2 = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n-2}.$$

The design of the test is based on the assumption of normality for ϵ_i, which necessitates that any linear function of Y_i will also be normally distributed, thus $\hat{\beta}_1$ will also follow a normal distribution. The assumptions of linear regression will be briefly discussed in Section 4.2.5.

One very common way of describing how much information the independent variables add to the model is the coefficient of determination, usually denoted R^2. It is a number between zero and one (closer to one is preferable). R^2 is defined as the fraction between the regression sum of squares and the total sum of squares:

$$R^2 = \frac{SS(\text{Regr})}{SS(\text{Total}_{\text{corr}})}.$$

The usefulness of R^2 is diminished by the fact that it increases with every added independent variable, regardless of the real significance of the variable. Thus, an adjusted coefficient of determination, R^2_{adj}, is often used instead, that punishes larger models more than R^2 does. R^2 is adjusted by creating a ratio of mean squares instead of sums of squares,

$$R^2_{adj} = 1 - \frac{\frac{SS(\text{Residual})}{n-p+1}}{\frac{SS(\text{Total}_{\text{corr}})}{n-1}} = \frac{MS(\text{Residual})}{MS(\text{Total}_{\text{corr}})}$$

Two other often used criteria used when developing a model are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). The AIC was defined in Akaike (1974) and in the linear regression framework it is

$$AIC(p') = n\ln(SS(\text{Residual})_p) + 2p' - n\ln(n). \tag{3}$$

The first two terms of (3) work against each other; as the first increases with added independent variables, the second decreases as the model gets larger.

The AIC tends to indicate too large models, therefore the BIC is sometimes preferred. The BIC was defined in Schwartz (1978) and in the linear regression framework it is

$$BIC(p') = n\ln(SS(\text{Residual})_p) + \ln(n)p' - n\ln(n). \tag{4}$$

The two expressions, (3) and (4), are very similar, but the BIC punishes larger models more, sometimes indicating a smaller model than the AIC (Rawlings, Pantula, and Dickey 2001).

Any statistical test where the test statistic has an F-distribution under the null hypothesis is called an F-test. In regression, F-tests most commonly test the hypothesis that the underlying data is better described by the smaller of two nested models. Thus, the null hypothesis is that several of the β parameters in the model are zero, e.g. $\beta_2 = \beta_3 = 0$, where the smaller model is $Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i$ and the full model is $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i$ and $\epsilon_i \in N(0, \sigma^2)$. All the parameters of both the models are estimated and their residual sums of squares are calculated. Using these, the test statistic is defined as

$$F = \frac{\frac{SS(\text{Residual})_{\text{reduced}} - SS(\text{Residual})_{\text{full}}}{n-(p'-k)-(n-p')}}{\frac{SS(\text{Residual})_{\text{full}}}{n-p'}} = \frac{\frac{SS(\text{Residual})_{\text{reduced}} - SS(\text{Residual})_{\text{full}}}{k}}{\frac{SS(\text{Residual})_{\text{full}}}{n-p'}}.$$

If F is larger than $F_\alpha(k,\ n-p')$, H_0 can be rejected in favour of H_1: that

at least one of the k β parameters are $\neq 0$, at significance level α (Rawlings, Pantula, and Dickey 2001).

4.2.5 Regression diagnostics

This part of the regression overview is aimed at studying the underlying assumptions of linear regression and how the validity of the model may be tested. The basic assumptions are, as have been stated in Section 4.2.1, that the columns (except for the first column of ones) of $\mathbf{X}$ contain n known constants, each. The random errors ϵ are assumed to have a common variance σ^2, zero mean and to be pairwise independent. This, together with the design of the linear model (expressions (1) and (2)), necessitates that the dependent variable, $\mathbf{Y}$, also follows these assumptions. Usually, the random errors and the dependent variable are also assumed to be normally distributed, for the sake of confidence and prediction intervals and test of significance (Rawlings, Pantula, and Dickey 2001).

In time series data, the errors tend to be correlated over time, for natural reasons. When the assumption that the errors are pairwise independent is broken the least squares estimates are unbiased, but might no longer be the best possible. Violating this assumption could also cause the variance estimates to be biased, making confidence and prediction intervals either too wide or to narrow. Also, test of significance could become less trustworthy, depending on the nature of the correlations.

The errors can also have non-constant variance. This problem can be solved in many ways, depending on the way the variance of the errors is heterogenous, e.g. by different transformations of the dependent data. Heterogenous variance makes the idea of equal weighting of each data point in the least squares estimator less optimal, since not all data points contain the same amount of information under heteroscedasticity.

Normality is not required for the least-squares estimator, nor for estimating the variance. Insofar as the other assumptions are not violated, the least squares estimates of the parameters are the best linear unbiased estimates possible. Confidence and prediction intervals, on the other hand, need the normality assumption in order to be correct. Also, test of significance, such as F- and t-test require normality. It should be noted that the F-test is quite robust to departures from the normality assumption (see Tiku (1971)). For large n, the t-test is also approximately valid, despite non-normality (Rawlings, Pantula, and Dickey 2001).

Another possible major problem with least squares linear regression is called the collinearity problem. It arises when a linear combination of some

of the columns in $\mathbf{X}$ (i.e. one or more of the independent variables) equals one of the other columns. This is the same as saying that the information contained in one of the independent variables is superfluous. Two cases can arise: if a linear combination of some of the columns exactly equals one of the other columns, a unique least squares solution cannot be found. If, on the other hand, the $\mathbf{X}$ matrix is only nearly singular, a unique solution can be found. If this situation arises, the β parameter estimates become unstable and small variations in the independent variable-space can lead to very different regression parameters (Rawlings, Pantula, and Dickey 2001). One common way to measure the effects of collinearity is to compute the so called variance inflation factor (VIF). The VIF is defined as

$$VIF_j = \frac{1}{1 - R_j^2},$$

where R_j^2 is the coefficient of determination corresponding to a regression where the independent variable X_j is used as dependent variable on the other independent variables. Thus, there is oneVIF_j for each of the independent variables. The VIF_j is proportional to the variance of the β_j parameter. There are several recommendations for when the problems of collinearity become serious. Rawlings, Pantula, and Dickey (2001) recommends a higher bound for the VIF of 10.

Another problem area of linear regression is the existence of outliers and influential points. As will be shown in Section 5, the problems with outlying points in the data at hand in this project are slight. Thus, the theory regarding outlying and influential points will not be overly discussed here. One commonly used method to study the influence of individual points on the regression is called Cook's Distance, developed by R. Dennis Cook in Cook (1977). It measures the effect that removing one specific observation has on $\hat{\beta}$. Consequently, there are as many Cook's Distances as there are observations. It is defined as

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})'(\mathbf{X}'\mathbf{X})(\hat{\beta}_{(i)} - \hat{\beta})}{(p+1)s^2}.$$

Cook's Distance can also be interpreted as the Euclidean distance between the original regression hyperplane, $\hat{\mathbf{Y}}$, and the regression hyperplane for the regression without data point i, $\hat{\mathbf{Y}}_{(i)}$ (Rawlings, Pantula, and Dickey 2001).

4.2.6 Categorical variables

In contrast to quantitative variables, that describe e.g. volume, speed or temperature, qualitative (or categorical) variables contain information regarding

classes or types, such as sex, colour or location. To be able to handle these in a regression analysis special care needs to be taken.

There is usually no good way to translate categorical variables to numerical values, so that they can be used in the $\mathbf{X}$ matrix in the previously mentioned linear regression models. If, for example, the categories are *Apple, Banana, Orange*, there is no good way to represent them as numerical values.

One solution is to introduce dummy (or indicator) variables that take the values 0 or 1. Dummy variables to describe the fruit classification above could be

$$X_{\text{Apple}} = \begin{cases} 1, & if\ X_{\text{fruit}} = \text{Apple}, \\ 0, & otherwise \end{cases} \quad X_{\text{Banana}} = \begin{cases} 1, & if\ X_{\text{fruit}} = \text{Banana}, \\ 0, & otherwise \end{cases}$$

$$X_{\text{Orange}} = \begin{cases} 1, & if\ X_{\text{fruit}} = \text{Orange}, \\ 0, & otherwise. \end{cases}$$

Columns containing zeros or ones could then be introduced into a $\mathbf{X}$ matrix, indicating which rows in the data belonged to which fruit class. This introduces problems for models that only contain categorical variables, since the matrix $\mathbf{X}'\mathbf{X}$ is singular (see Section 4.2.5). This can be solved by removing a column (i.e. a fruit) or by removing the intercept. This kind of reparamertisation is still used, but after the advent of computers it has largely been replaced by the generalised inverse approach, which uses one of the non unique solutions to the normal equations, despite there being no one unique solution. Details regarding the two approaches can be found in Rawlings, Pantula, and Dickey (2001).

4.3 Time series analysis

The following subsections will outline time series modelling theory, focusing on the classical AR, MA and ARMA models, that all let the current data point depend linearly on preceding data points, as well as some extensions to these.

4.3.1 AR, MA, ARMA and ARIMA models

The moving average (MA(q)) process of order q is defined in Jakobsson (2013) p. 58 and 62 as

Definition 1.

$$y_t = e_t + c_1 e_{t-1} + \ldots + c_q e_{t-q} = C(z)e_t,$$

where $C(z)$ is a monic polynomial of order q, i.e.,

$$C(z) = 1 + c_1 z^{-1} + c_2 z^{-2} + \ldots + c_q z^{-q},$$

where $c_q \neq 0$, e_t is a zero-mean white noise process with constant variance σ^2 and z^{-1} is the unit delay operator, defined as,

$$z^{-1}x_t = x_{t-1}.$$

As can be seen from Definition 1, the MA process can be interpreted as a linear regression with the current time series value as dependent variable and both the q previous and current white noise terms as independent variables.

The second common linear process is the atoregressive (AR(p)) process of order p. It describes time-varying processes where the output variable depends linearly on its own previous values. It is defined in Jakobsson (2013), p. 67, as

Definition 2.

$$A(z)y_t = y_t + a_1 y_{t-1} + \ldots + a_p y_{(t-p)} = e_t,$$

where $A(z)$ is a monic polynomial of order p, i.e.,

$$A(z) = 1 + a_1 z^{-1} + \ldots + a_p z^{-p},$$

where $a_p \neq 0$ and e_t is a zero-mean white noise process with constant variance σ^2.

The two processes defined above can be merged into the autoregressive moving average (ARMA($p,\ q$)) process, defined in Jakobsson (2013), p. 77, as

Definition 3.

$$A(z)y_t = C(z)e_t,$$

where $A(z)$ and $C(z)$ are the monic polynomials defined in Definitons 2 and 1, respectively. As in the previous definitions, e_t is a zero-mean white noise process with variance σ^2.

Finally, the autoregressive integrated moving average (ARIMA($p,\ d,\ q$)) model is used in the case when the data to be modelled is not stationary but instead exhibits a trend. It is defined in Jakobsson (2013), p. 112, as

Definition 4.

$$A(z)(1 - z^{-1})^d y_t = C(z)e_t,$$

where $A(z)$, e_t and $C(z)$ are defined as previously and d denotes the number of differentiations.

The most common value for d is one, and only very seldom is it above two. In practice, ARIMA models are used by first forming a new, differentiated time series, which is then used to build an ARMA model.

For details regarding e.g. estimation of the process parameters, prediction, or other parts of the theory, see Jakobsson (2013) or other introductions to time series modelling.

4.3.2 Model structure identification

Identifying a times series model structure is not a straightforward process. Generally, it involves choosing a model structure and order, validating that choice and then returning to change the structure/order again, until acceptable results are attained (Jakobsson 2013).

Two common tools used for modelling are the auto-covariance function (ACF), $\rho(k)$, and the partial auto-correlation function (PACF), $\phi(k)$. Together they can indicate a reasonable initial model for the time series in question. An MA(q) process should have a decaying sine and/or exponential behaviour in the PACF, and $\rho(k) = 0$ for lags greater than the MA order, q. An AR(p) process exhibits a decaying sine and/or exponential function in the ACF, instead, and $\phi(k) = 0$ for lags greater than the AR order of the process, p. Finally, an ARMA(p, q) process will have a damped sine and/or exponential behaviour in both the ACF and PACF, as is summarised in a table on p.101 in Jakobsson (2013).

ARIMA(p, d, q) models are used when the process to be modelled is not stationary, i.e. when it has some kind of trend. Indications of this can be seen in the ACF, as it decays very slowly (Jakobsson 2013).

4.3.3 Non-constant variance - GARCH models

The autoregressive conditional heteroscedastic (ARCH) models were introduced in Engle (1982). ARCH models are used to model processes whose conditional variances change as a function of past squared values of the process. The ARCH(q) model is defined as

Definition 5.

$$\begin{aligned} y_t =& \sigma_t \epsilon_t, \\ \sigma_t^2 =& \omega + \sum_{i=1}^{q} \alpha_i y_{t-i}^2, \end{aligned}$$

where $\alpha_i > 0,\ \omega > 0,\ \sum_{i=1}^{q} \alpha_i < 1$ *and the* ϵ_t *are a sequence of zero mean unit variance i.i.d. r.v..*

As the previously defined time series models, the ARCH model can be seen as a regression model. The ARCH model was expanded into the generalised ARCH (GARCH) model in Bollerslev (1985). It adds a sum of conditional variances to the model in Definition 5,

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i y_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2.$$

This can be rewritten as,

$$\beta(z)\sigma_t^2 = \omega + \alpha(z) y_t^2,$$

where $\beta(z) = 1 - \beta_1 z^{-1} - \ldots - \beta_p z^{-p}$ and $\alpha(z) = \alpha z^{-1} + \ldots + \alpha_q z^{-q}$. We see that a GARCH model can be described as assuming an ARMA model for the error variance.

The GARCH model for the error variance can be combined with an ARMA model for the mean structure as (Cryer and Chan 2008)

$$\begin{aligned} A(z) y_t &= C(z)\epsilon_t, \\ \epsilon_t &= \sigma_t \varepsilon_t, \\ \beta(z)\sigma_t^2 =& \omega + \alpha(z)\epsilon_t^2. \end{aligned}$$

ARMA-GARCH models are built by first modelling the ARMA-part based on y_t and then using the residuals from the ARMA model to model the GARCH-part. The order of a GARCH model can be decided in much the same way as an ARMA model, by looking at ACF and PACF of the squared residuals (Cryer and Chan 2008). The presence of ARCH/GARCH effects can also be studied by using a so called McLeod-Li test, described in McLeod and Li (1983). (Baum 2013)

4.3.4 Long term memory - ARFIMA models

Long term memory is a term used to describe time series that exhibit dependence between temporally distant observations. Such dependence has been shown to exist in various time series, coming from many applications,

such as meteorology, finance and hydrology (Vasilev 2007). In the time domain, a tong term memory time series is characterised by the fact that its autocorrelation decays very slowly.

Non-stationary time series are modelled using ARIMA models, as defined in (4). There are several tests for non-stationarity. One common such is the Elliott-Rothenberg-Stock test, which has non stationarity as null hypothesis. Often, time series for which unit root tests, such as the Elliott-Rothenberg-Stock test, indicate non stationarity are modelled using an ARIMA model (Baum 2013). Unit root tests can be complemented by test that have stationarity as null hypothesis, such as the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test. Using two such tests together makes it possible to discern if a time series is more likely to be best described as non stationary (has a unit root, I(1), stationary I(0), or if it might exhibit long memory, since long memory series are neither I(1) nor I(0) (Baum 2013).

ARFIMA models can be defined in much the same way as ARIMA models where in Definition 4., with the difference that the order of the differencing, d, is allowed not only to take integer values, but also fractional ones. Then, $(1 - z^{-1})^d$ is redefined as the fractional differencing operator

$$(1 - z^{-1})^d = \sum_{k=0}^{\infty} \frac{\Gamma(k - d)z^{-k}}{\Gamma(-d)\Gamma(k + 1)},$$

where $\Gamma(\cdot)$ denotes the generalised factorial function (Baum and Wiggins 2000). The value of d describes the behaviour of the process. A process with $d \in (-0.5, 0)$ demonstrates negative long memory, a process with $d \in (0, 0.5)$ demonstrates long term memory, a process with $d = 0$ is a stationary ARMA process and a process with $d \in (0.5, 1)$ exhibits no long term temporal dependence. A process with $d = 1$ is, of course, an ARIMA(p, 1, q) process (Baum and Wiggins 2000).

4.3.5 Tidal harmonic analysis

As is briefly described in Section 2.4, oceanographic data can be partly described by deterministic tidal components. That is, they can be divided into a tidal signal and a non-deterministic noise signal. One method for estimating the components of the tidal signal is classical harmonic analysis, in which the tide is defined as the sum of specific sinusoids, whose frequencies are related to different astronomical parameters. There are six fundamental frequencies which are used to specify the frequencies of the sinusoids. They are: the rotation of the earth (24.8 h), the orbit of the moon (27 days), the orbit of the earth (tropical year), the lunar perigee (8.85 years), lunar orbit

tilt (18.6 years) and the parihelion (21000 years) (Pawlowicz, Beardsley, and Lentz 2002).

Arthur Thomas Doodson was a British oceanographer who designed a practical system for specifying the different harmonic components of the tide, in Doodson (1921). The system makes use of Doodson numbers, which is a six digit number that can be used to describe every tidal constituent. The tidal constituents are named according to a system developed by George Darwin. For example, the principal lunar semidiurnal constituent is named M_2 and the solar annual constituent is S_a. In classical harmonic tidal analysis, least squares fitting is used to estimate the phase and amplitude of each frequency (Pawlowicz, Beardsley, and Lentz 2002).

The tidal response is modelled as

$$x(t) = b_0 + b_1 t + \sum_{k=1,\ldots,N} A_k \cos(\sigma_k t) + B_k \sin(\sigma_k t). \tag{5}$$

Here, there are N tidal constituents (with a unique Doodson number each), with known frequencies and unknown amplitudes. The first two terms in (5) are optional in the package used in this project and handle offset and drift.

One possible problem with the classical harmonic analysis method is that the tidal response in coastal regions can be influenced by such factors as geography and water salinity. Also, if the water depth is shallow compared to the tidal wave height, non-linear effects can be present. Problems such as these can be partly remedied by including so called shallow-water constituents in the harmonic analysis. The MATLAB package used in this projects makes use of a maximum of 45 astronomical and 101 shallow-water constituents. This package also improves the classical harmonic tidal analysis method by calculating confidence intervals for the analysed components. Details about how these confidence intervals are estimated, as well as more theory regarding the harmonic analysis can be found in Pawlowicz, Beardsley, and Lentz (2002).

5 Results

The results section is divided into three main parts - extreme value analysis (Section 5.1), regression analysis (5.2) and time series analysis (5.3). Some results are also presented in the Appendix, especially when the results for different locations are very similar. As a rule of thumb, the Uddevalla results will be the results that are primarily presented in the text.

5.1 Extreme value analysis

As is discussed briefly in the Introduction (Section 1), the on duty oceanographers at SMHI know through experience that certain locations along the Swedish coast sometimes exhibit more extreme sea level behaviours than other locations. Three such locations are the locations of the mobile sea level gauges at Uddevalla, Ängelholm and Åhus. One way of motivating their placement, i.e. studying if the sea level behaviour merits placing extra gauges, is to study the distribution of the differences between the mobile gauges and their paired permanent stations: Smögen, Viken and Simrishamn.

When comparing two different gauging locations, a natural goal is to be able to say how big the difference between the sea levels at the two positions can become on a long time scale. To this effect, extreme value theory can be used. The method used in this project goes as follows: divide the complete data set into equally sized blocks, create a new time series of the block maxima, fit a GEV distribution to the block maxima and evaluate the fit.

As is discussed in Section 4.1, extreme value analysis is useful for describing the behaviour of extremes of data sets. Here, the differences between the data from the mobile gauges and their permanent counterparts will be studied. In order for the assumptions of the GEV distribution to hold, the data need to be independent. Thus, the data need to be separated into blocks of such sizes that the block maxima will form a time series that is independent enough for the GEV distribution to fit well and the extreme value analysis to work. When choosing the block sizes, there is a trade-off between bias and variance (Coles 2001). Choosing few data points (maxima of large blocks) can increase the variance of parameter estimates and thus make extrapolation very uncertain. On the other hand, choosing to include many data points (maxima of small blocks) may induce bias, as the fit of the data to the GEV model may be bad due to dependence in the underlying data.

In order to find a good block size two methods are used. To check the

independence of the block maxima for specific block sizes autocorrelation plots are used, since they can be utilised to examine the similarity of the data to a white process (Jakobsson 2013). To check how the variance of the estimates change with block size, plots of five-year return levels together with 95 % confidence intervals are plotted. The return levels and their confidence intervals are calculated as described in Section 4.1.2.

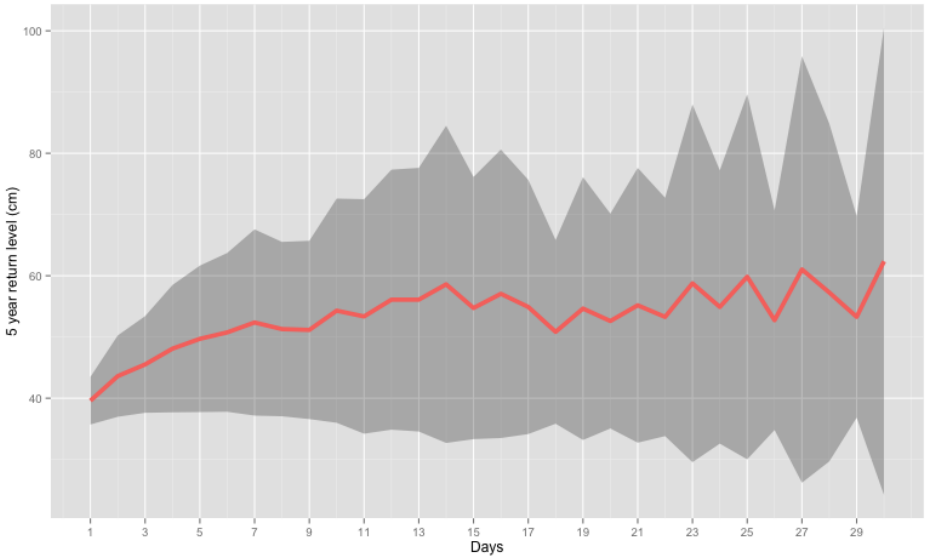

Figure 9: The five year return levels corresponding to GEV distributions fitted to block maxima of the difference between the Uddevalla measurements and the measurements from Smögen are shown by the red line. The block sizes are on the x axis, in multiples of 24 hours. A 95 % confidence band for the return levels is also shown, calculated by the delta method.

Five-year return levels for the difference between the Uddevalla and Smögen time series are shown together with their corresponding confidence intervals in Figure 9. The return levels appear to converge to approximately 55 cm and for block sizes above 10 days they stay roughly constant, indicating that the assumptions are met well enough to avoid bias. For block sizes above 22 days, the variance increases, as is expected considering the aforementioned trade-off between bias and variance.

In Figure 10 the autocorrelation of the block maxima series for one specific block size is shown. The data is divided into 65 two-week blocks. The block size choice is based on the results presented in Figure 9. The ACF further implies that the data used to fit the GEV distribution is stationary.

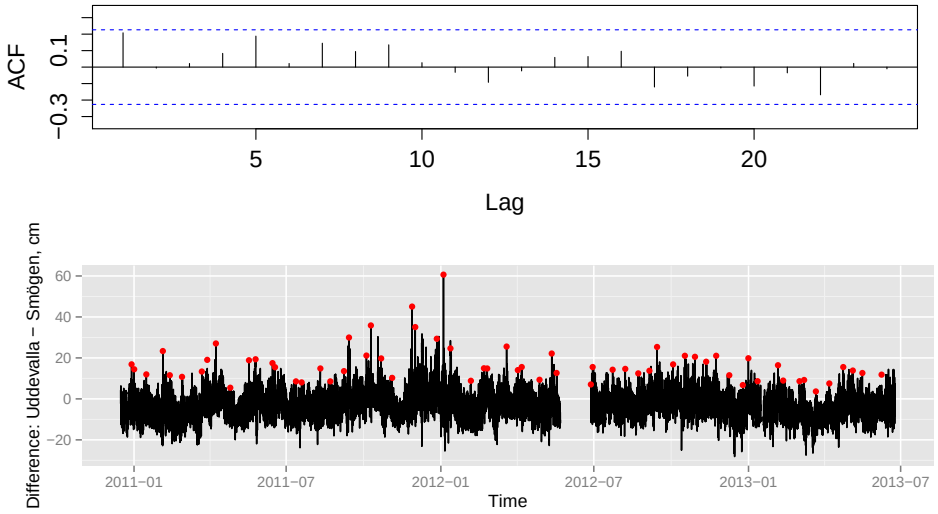

Figure 10: Top: an autocorrelation plot for the Uddevalla - Smögen two-week block maxima series. Bottom: a plot showing the 65 block maxima (red dots) over the difference data.

The GEV distribution fit is diagnosed in Figure 11. Four diagnostic plots are presented: a probability plot, a quantile-quantile plot, a density plot and a return level plot. They all imply a good GEV distribution fit for the two-week block maxima series. The GEV distribution for this block size has a location parameter μ that is 12.6337 (standard error 0.81764), a scale parameter σ that is 5.8407 (standard error 0.62377) and a shape parameter ξ that is 0.1375 (standard error 0.09111). The positive value of the shape parameter implies a concave return level function. This is not physically reasonable and therefore using the GEV fit to calculate return levels for very long return periods might not be wise.

For Uddevalla and Smögen, the most extreme differences can be found in the cases where the sea levels are higher in Uddevalla than in Smögen. For the sake of completeness, the inverted case is also studied, i.e. the extremes of the Smögen - Uddevalla data. This is the same as studying the minima of the difference between Uddevalla and Smögen. Plots corresponding to Figures 9 – 11 are presented for this case in Appendix B.1.

The same analysis has also been made for the differences between the sea levels in Ängelholm – Viken and Åhus – Simrishamn. The results for these

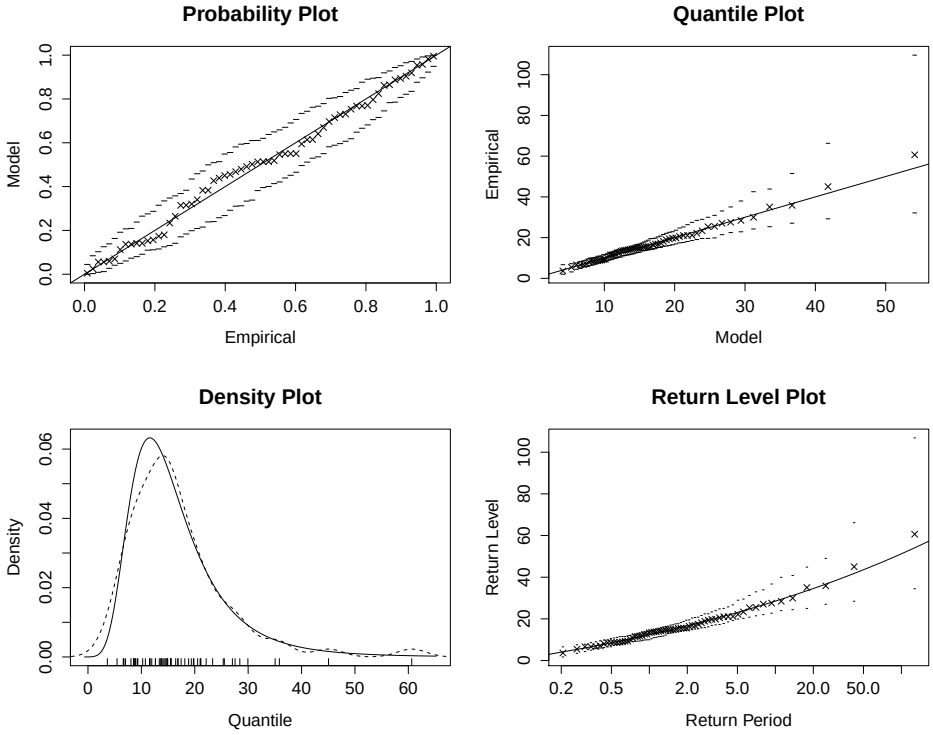

Figure 11: Diagnostic plots for the GEV fit to the Uddevalla - Smögen difference data. Left-to-right and top-to-bottom: 1) A probability plot. The data points follow a straight line reasonably well, considering the 95% confidence band. 2) A quantile-quantile plot further implies that the fitted model suits the data, since the model-based estimate of the quantile function fits the data well, considering the confidence bands. 3) The density of the fitted distribution is plotted (solid line), together with a non-parametric estimate (dashed line) and a rug plot of the data. The fit is quite good, but has to low an upper tail to account for the largest block maximum. 4) The return level plot also shows that the fit is reasonably good, since the data points lie close to the return level line. The x-axis shows return levels (in multiples of two weeks) plotted on a logarithmic scale. The line is concave (and thus has no finite bound), which corresponds to the positive value of ξ in the GEV-fit.

locations are plotted in the same manner as the Uddevalla – Smögen results in Appendices B.2 and B.3.

The different five year return levels are presented together with 95 % confidence intervals in Table 3. As would be expected, the largest return

level is found for the Uddevalla – Smögen case - just over half a meter. Also not entirely unexpected, considering the bad GEV fit to the Simrishamn – Åhus data presented in Appendix B.3, the widest confidence interval by far is the one for the Simrishamn – Åhus case.

	Lower	z_p	Upper
Uddevalla – Smögen	34.04	53.10	73.67
Smögen – Uddevalla	24.51	29.12	33.73
Ängelholm – Viken	23.76	35.38	47.00
Viken – Ängelholm	22.89	38.36	47.70
Åhus – Simrishamn	18.35	25.34	32.26
Simrishamn – Åhus	6.65	49.11	91.56

Table 3: Five year return levels for the different cases studied in this Section (and in Appendices B.1 to B.3). Also, the upper and lower points of their respective 95 % confidence intervals are given.

5.2 Regression analysis

Now that the differences between the mobile gauges and their permanent counterparts have been studied, an attempt is made at relating the sea levels at the mobile gauges with the sea level data from the permanent gauges, according to the interests of SMHI, as stated in the Introduction, Section 1.

The relations between the data sets will be studied through regression analysis. Initially, simple linear regression models are designed, in Section 5.2.1. These regression models are diagnosed and extended into multiple linear regression models in Sections 5.2.2 – 5.2.4. In the same manner as for the extreme value result, the regression analysis results from Uddevalla – Smögen are primarily presented in the text while the results from Ängelholm – Viken and Åhus – Simrishamn/Kungsholmsfort are partly presented in Appendix C. Throughout this section (and Appendix C) the data is divided into parametrisation and validation data. The parametrisation data make up three-quarters of the complete data set and the validation data make up the remaining quarter.

5.2.1 Initial linear regression

To relate the sea levels in the paired locations, linear regression models are used, according to the theory presented in Section 4.2.1. Four simple linear

regression models are built, one each for Uddevalla – Smögen and Ängelholm – Viken and two for Åhus (Åhus – Simrishamn and Åhus – Kungsholmsfort). They are summarised in Table 4. The β-parameter for the Uddevalla–Smögen regression is clearly larger than the other three, corresponding to a steeper regression line. Also, both the BIC and adjusted R^2 indicate that the Simrishamn data set is preferred as independent variable for the Åhus regression. All four regressions have very high adjusted R^2 values, as might be expected, considering the correlations given in Section 3.1.3 and the clear linear relationships shown in Figure 7.

	$\hat{\alpha}$ (S.E.):	$\hat{\beta}$ (S.E.):	Adj. R^2:	BIC:
Uddevalla – Smögen	−2.9913	1.1269	0.9423	104226
	(0.0522)	(0.0022)		
Ängelholm – Viken	−3.1544	1.0730	0.9634	78185.75
	(0.0380)	(0.0018)		
Åhus – Simrishamn	3.6451	1.0165	0.9829	46505.6
	(0.0336)	(0.0014)		
Åhus – Kungsholmsfort	−0.2545	1.0243	0.9533	56057.13
	(0.0515)	(0.0023)		

Table 4: Regression parameter values for the four initial linear regressions together with their standard errors. The adjusted R^2 and Bayesian Information Criteria values are also given. The model parametrisation data points are included in this regression.

Figure 12 shows the linear regression line $\hat{Y}_{i,\text{Udd.}} = \hat{\alpha} + \hat{\beta} X_{i,\text{Smö.}} = -2.99 + 1.13 X_{i,\text{Smö.}}$. 95.7 % of the validation data points fall within the theoretical prediction band, as is shown in the figure. For Ängelholm and Åhus (with Simrishamn as independent variable) the same quotient is 95.4 % and 95.1 %, respectively. The root mean square errors (RMSEs) of the four linear models are for Uddevalla: 6.45 cm, for Ängelholm: 4.29 cm and for Åhus: 2.80 cm (Simrishamn) or 5.00 cm (Kungsholmsfort). Thus, the RMSE for the validation data also indicates that the Simrishamn data is better as independent variable in the Åhus regression.

5.2.2 Initial regression diagnostics

Diagnosing a regression model can be done in many ways. Here, some diagnostic plots and methods are used, though there are many more. Diagnostic plots for the Uddevalla – Smögen regression are shown in Figure 13. Starting

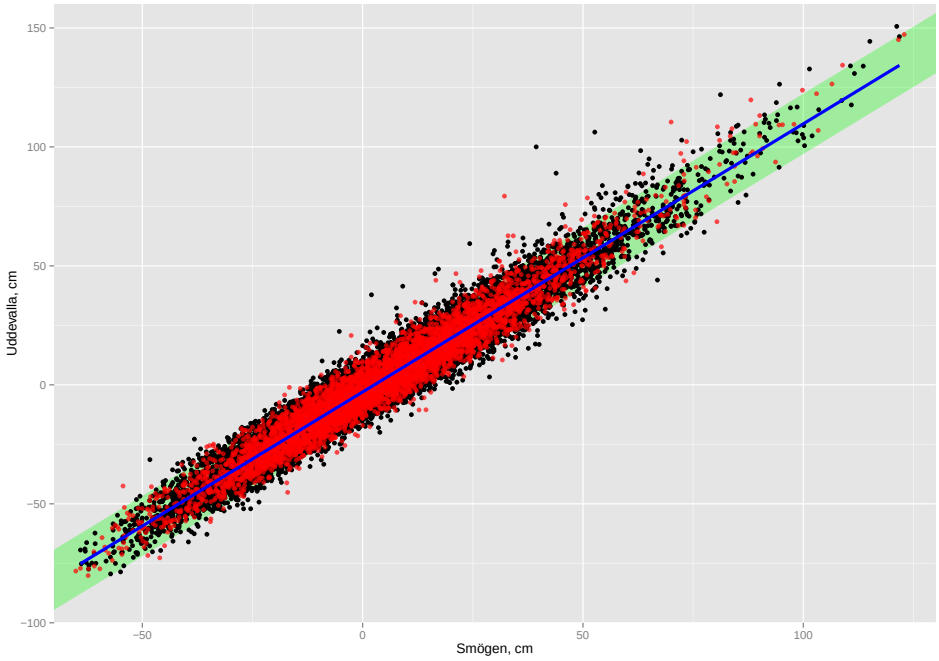

Figure 12: A linear regression line between the Uddevalla and Smögen data is shown in blue. It has, $\alpha = -2.991$ (0.052), $\beta = 1.127(0.002)$, with standard errors in the brackets. The line was calculated using the data point shown in black. The red points indicate validation data. 95.7 % of the validation points are within the theoretical 95 % prediction band, shown in green.

with the quantile-quantile plot, it shows that the normality assumption is not met completely, as there are some points quite clearly diverging from the straight line, indicating a distribution with a thicker upper tail than the normal distribution. Considering the effects of non-normality on the regression, as discussed in Section 4.2.5, this violation of the regression assumptions is not regarded as problematic, especially since the prediction intervals are very nearly correct (95.7 % instead of 95 %).

The Cook's distance plot does not indicate any large problems with regard to influential points. The largest Cook's distance is 0.01, indicating that the change in the regression line, $\hat{Y}$, would be small if the most influential point is removed.

Two other aspects stand out when analysing the diagnostic plots in Figure 13. The first is potentially problematic: all three plots indicate the same three data points as being the most extreme. They are both outliers (in the residual) and the most influential according to the Cook's distance. Closer

inspection of the Y- and X-values that correspond to these residual outliers show that they are outliers in neither the Y- nor X-spaces. Furthermore, they are not obviously faulty measurements, as the points indicated in Section 3.1.1 are. Thus, they are left in the data set and may hopefully be better explained as the initial regression model is extended, in the following sections.

The second aspects that stands out is the (slight) quadratic trend in the plots of the regression residuals e_i versus the fitted values, $\hat{Y}_i$. This will be dealt with in the following section.

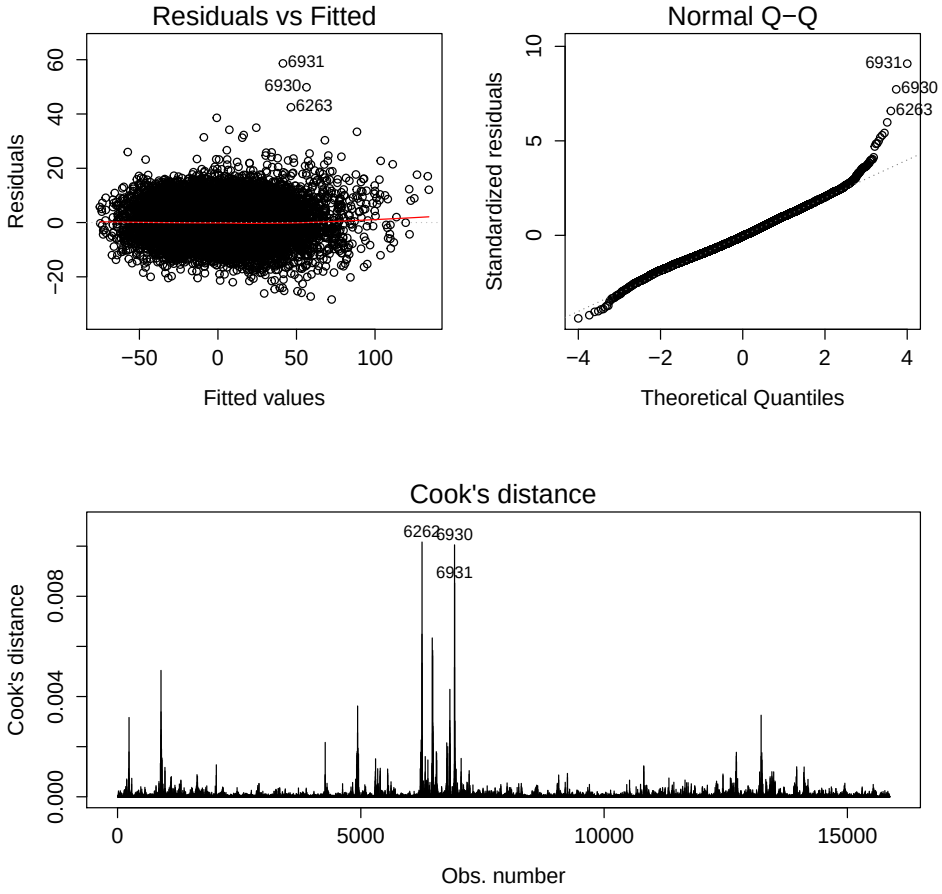

Figure 13: Top-to-bottom and left-to-right: a plot of regression residuals vs. the fitted values, a normal quantile-quantile plot, and a plot showing Cook's distance.

The diagnostic plots for the other two regressions are shown in Appendix

C. The same slight curve is visible in their plots regression residuals e_i versus the fitted values, $\hat{Y}_i$ as it is in Figure 13.

One other obvious problem with these regressions is the temporal dependence between the errors. Correlated errors can bias the variance estimates, making confidence and prediction intervals incorrect, as mentioned in Section 4.2.5. This appears to not be the case, as the prediction intervals fit the validation data quite well, for all three regression models. Nonetheless, this calls the validity of the regression model into question, since the regression assumptions are not met.

5.2.3 Extension of the initial linear regression

The initial linear regressions are extended to polynomial regression models, as described in Section 4.2.2. This is done in order to compensate for the (slight) curve visible in the plots of the regression residuals e_i versus the fitted values, $\hat{Y}_i$ for all three models. The slightly curved structure of the residuals implies that there is some structure in the data that is not captured by the initial linear regression. Different polynomial regression models are tested. These models, up to the third order, are summarised in Table 5. Also, BIC values are given. These, together with two-way F-tests are used to determine what model fits the given data best, in all three cases. In the Uddevalla–Smögen case, the best fitting model is $y = \alpha + \beta_1 x + \beta_3 x^3$. In the Ängelholm–Viken case, the best fitting model is $y = \alpha + \beta_1 x + \beta_2 x^2$. Finally, in the Åhus-Simrishamn case, it is also $y = \alpha + \beta_1 x + \beta_2 x^2$. The Uddevalla case necessitates some more attention. Considering the recommendation in e.g. Faraway (2004) to never remove a lower order term in a model, in order to avoid adding additional terms to the model under a scale change, the $y = \alpha + \beta_1 x + \beta_3 x^3$ model is discarded. Also, considering the very small differences between the two remaining models, the more parsimonious model $y = \alpha + \beta_1 x + \beta_2 x^2$ is chosen. Looking at plots of the residuals versus fitted values for these models, the curved structure that is visible in the plots for the smaller models is no longer evident.

Figure 14 shows the difference between the initial linear regression model, and the model including a quadratic term, for the Uddevalla – Smögen regression. The difference is very slight, except for the highest and lowest X-values. The regression line confidence intervals do not overlap, for X values over 70 cm.

Another possible extension is to include both the data from Simrishamn and Kungsholmsfort in the Åhus regression: $Y_{i,\,\text{Åh.}} = \alpha + \beta_1 X_{i,\,\text{Sim.}} + \beta_2 X_{i,\,\text{Kung.}} + \epsilon_i$. This increases the $R^2_{\text{Adj.}}$, but the variance inflation factor is

	$\hat{\alpha}$ (S.E.):	$\hat{\beta}_1$ (S.E.):	$\hat{\beta}_2$ (S.E.):	$\hat{\beta}_3$ (S.E.):	BIC:
U–S	−3.1280 (0.0595)	1.1217 (2.451e-03)	2.880e-04 (6.057e-05)	–	104213.1
	−3.0067 (0.0522)	1.1146 (2.981e-03)	–	5.779e-06 (9.347e-07)	104197.5
	−3.0289 (0.0645)	1.1148 (3.013e-03)	4.976e-05 (8.500e-05)	5.250e-06 (1.312e-06)	104206.8
Ä–V	−3.3637 (0.0416)	1.0690 (1.824e-03)	4.625e-04 (3.879e-05)	–	78053.78
	−3.1694 (0.0380)	1.0664 (2.191e-03)	–	2.461e-06 (4.688e-07)	78167.73
	−3.3771 (0.0245)	1.0711 (2.222e-03)	5.044e-04 (4.663e-05)	-9.086e-07 (5.612e-07)	78060.67
Å–S	3.8141 (0.0357)	1.0070 (0.0015)	4.9127e-04 (3.710e-05)	–	46340.96
	3.6040 (0.0345)	1.0088 (2.039e-03)	–	3.750e-06 (7.294e-07)	46488.36
	3.8273 (0.0388)	1.0082 (2.024e-03)	-5.074e-04 (4.153e-05)	-6.9996e-07 (8.103e-07)	46349.38

Table 5: For all three pairs of stations, three different polynomial models have been fitted. They are of the form $y = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$. Each row of the table corresponds to a model.

24.15, indicating that collinearity has a large effect. This is not surprising, since the linear correlation between the Simrishamn and Kungsholmsfort data is 0.98. Thus, to keep the β parameter estimates stable, this regression model is discarded.

5.2.4 Multiple linear regression

Wishing to see whether the wind and atmospheric pressure data can make a significant difference to the previous regression, a multiple regression model is designed. All of the different models presented in this section have been designed using the training data, as previously discussed. As stated in Section 3.2, the directional wind data is given in degrees. Thus, in order to

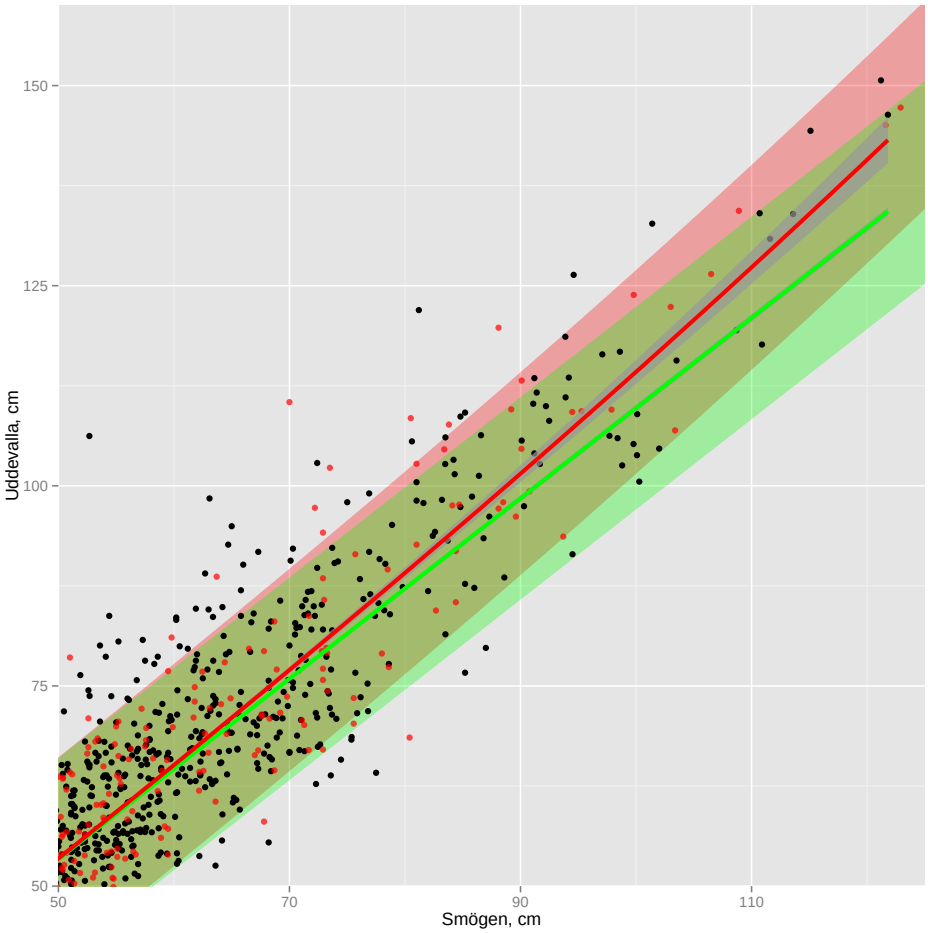

Figure 14: Two different linear regression lines for Uddevalla-Smögen are plotted together. In red: $y = \alpha + \beta_1 x + \beta_2 x^2$. In green: $y = \alpha + \beta_1 x$. The semi-transparent red and green bands are 95% prediction intervals and the gray bands are 95% confidence bands for the regression lines. As in previous plots, the black points are training data and the red ones are validation.

capture the directional nature of the data, the wind data is structured into eight directions, as shown in Figure 8. Initially, a model without interactions is designed. Using BIC and stepwise F-tests, the following model is found to be the best fit for the Uddevalla data:

$$y_{\text{Udd}} = \alpha + \beta_1 x_{\text{Smö}} + \beta_2 x_{\text{Smö}}^2 + \beta_3 x_{\text{ws}} + \beta_4 x_{\text{NE}} + \beta_5 x_{\text{E}} + \beta_6 x_{\text{SE}} + \beta_7 x_{\text{S}} + \beta_8 x_{\text{W}}. \tag{6}$$

Here, $x_{\text{NE}}, \ldots, x_{\text{W}}$ correspond to dummy-variables that represent the eight different directions and x_{ws} is the wind speed. All of the included coefficients are significant at a 0.0001% significance level. The signs of the regression coefficients show that both the Smögen coefficient as well as the squares Smögen parameter have a positive effect on the sea level in Uddevalla. The same goes for the westerly winds. The winds from NE, E, SE and S, on the other hand, all have a negative effect on the sea levels in Uddevalla. This makes a lot of sense considering the geography of the Skagerrak coastline, as discussed in Section 2.4. Observe that some wind directions as well as the pressure data have been discarded as not significant, both when using BIC and F-tests. The BIC value of the best Uddevalla model is 103483.2 and the adjusted R^2 is 0.9452. The RMSE on the validation data set is 6.26 cm. The largest of the variance inflation factors in this model is 1.60, indicating low levels of collinearity between the independent variables.

Using the same notation, the best fitting model for the Ängelholm data is:

$$\begin{aligned} y_{\text{Äng}} =& \alpha + \beta_1 x_{\text{Vik}} + \beta_2 x_{\text{Vik}}^2 + \beta_3 x_{\text{ws}} + \beta_4 x_{\text{N}} + \beta_5 x_{\text{NE}} + \beta_6 x_{\text{E}} + \beta_7 x_{\text{SE}} + \beta_8 x_{\text{S}} \\ &+ \beta_9 x_{SW} + \beta_{10} x_W + \beta_9 x_{\text{pr.}}. \end{aligned}$$

Here, the same notation is used as in (6), but the pressure data has been added, $x_{\text{pr.}}$. Again, all of the included coefficients are significant at a 0.0001% significance level. As for Uddevalla, the β-parameters corresponding to the data from Viken, and their squares, are positive, as are the wind speed, SW and W parameters. The parameter corresponding to N, NE, E, SE, S and atmospheric pressure are negative. Looking at the map in Figure 1, this makes sense as winds pushing water into Skälderviken from west should make the sea levels in Ängelholm harbour rise, while winds pushing water out of the bay should make the sea levels fall. As was mentioned in Section 2.4, atmospheric pressure tends to lower sea levels. The BIC value of the best Ängelholm model is 76010.36 and the adjusted R^2 is 0.9691. The RMSE on the validation data set is 3.95 cm. The largest of the variance inflation factors in this model is 2.44, indicating low levels of collinearity between the independent variables, though not as low as for the Uddevalla regression.

For Åhus, the best fitting model is:

$$\begin{aligned} y_{\text{Åh}} = \alpha &+ \beta_1 x_{\text{Sim}} + \beta_2 x_{\text{Sim}}^2 + \beta_3 x_{\text{ws}} + \beta_4 x_{\text{N}} + \beta_5 x_{\text{E}} + \beta_6 x_{\text{SE}} + \beta_7 x_{\text{S}} \\ &+ \beta_8 x_{\text{SW}} + \beta_9 x_{\text{pr.}}. \end{aligned}$$

Again, all of the included coefficients are significant at a 0.0001% significance level. Here, the parameters corresponding to the sea water levels in Simrishamn, E, SE, S and SW are positive while the squares of the Simrishamn sea levels, the winds speed, N and the atmospheric pressure parameters are negative. Again, looking at the map in Figure 1, this makes sense as winds pushing water into the part of Hanöbukten where Åhus harbour is located from the east, south-east and south should make the sea levels in Åhus harbour rise. The BIC value of the best Åhus model is 45346.52 and the adjusted R^2 is 0.9849. The RMSE on the validation data set is 2.62 cm. All of the variance inflation factors in this model are below 1.4, indicating low levels of collinearity between the independent variables.

The parameter values for all of these regression models are presented in Appendix C, Table 9.

Fo two of these models, Uddevalla and Ängelholm, the β-parameters for wind speed are positive, while it is negative for the Åhus model. It makes sense that the interactions between wind speeds and wind directions might better explain the sea levels, than wind speeds and directions do separately. Models with interactions can be estimated using the same methods as models without interactions. They may add some predictive power, but interpretation of the parameters is not a straightforward and the models are much less parsimonious. This, together with the generally high R^2 values of the models presented so far indicate that they might not add much to the current models. The problems with interaction models are discussed in e.g. Faraway (2004).

5.3 Time series analysis

As is mentioned in Section 5.2, the temporal dependence in the regression errors casts the validity of a regression model into some doubt. One possible next step is to analyse the data with time series analysis. That is, build AR/MA/ARMA... models to describe the behaviour of the time series in question. In Section 5.3.1, initial time series models are designed for the three mobile sea level gauge data sets. They are reexamined in 5.3.3 after deterministic tidal structures in the data are studied in 5.3.2. Also, time series models are built for the data from the paired permanent stations, after which the joint behaviour of the time series model residuals are studied.

5.3.1 Initial models

Using the tests and visual methods described in Section 4.3, ARFIMA-GARCH models are built for the three data sets from the mobile gauges. As

previously, the results for Uddevalla are presented in the main text, while the results for Ängelholm and Åhus are mainly presented in the appendix.

Initially, a parsimonious model is sought after to describe the mean behaviour of the Uddevalla data set. Unfortunately, no such model could be found. Since both tests as well as information criteria and visual inspections of ACF and PACF for the residuals indicated that a fractionally integrated model is preferred to a model with an integer valued integration for the data, an ARFIMA($p,\ d,\ q$) model is chosen. The McLeod-Li test, as well as visual inspections of the ACF of the squared standardised residuals indicate the presence of ARCH/GARCH effects. Thus, an ARFIMA-GARCH model structure is chosen.

Using visual inspections of the ACF and PACF as well as the BIC, an ARFIMA-GARCH(27,d,10)-(1,2) model is chosen for the Uddevalla data. Most of the ARMA parameters in this model are set to zero. The parameters that are included and estimated are: $a_1 - a_5,\ a_{11} - a_{13},\ a_{18},\ a_{24},\ a_{26},\ a_{27},\ c_4,\ c_5,\ c_8 - c_{10}$. Thus, 12 ARMA parameters are included, together with four GARCH parameters and a degree d for the fractional integration. The residuals exhibit more kurtosis (heavier tails) than the normal distribution, thus a student's t-distribution is chosen for the model innovations. The fractional integration order, d, is estimated to 0.15. This indicates the presence of long memory. Despite the large number of parameters, not all structure in the data is accounted for by the model, as is shown in Figure 15.

The empirical density and qq-plots show that the student's t-distribution is a good choice for the data at hand. The ACF-plot of the squared residuals does not indicate problems, either, indicating that the GARCH part of the model succeeds at modelling the variance structure. The ACF-plot for the standardised residuals, on the other hand, shows that there is still undescribed structure left in the data. It should be noted that the autocorrelation for some lags is large compared to the confidence intervals, but still quite small, considering the scale on the y-axis.

The choice to halt the size of the model here is largely due to two facts. The first is that the optimisation algorithm fails to converge when certain parameters are introduced to the model, especially MA-parameters. The second is parsimony – smaller models might be fitted if the data are preprocessed.

5.3.2 Tidal harmonic analysis

It is clear from Section 5.3.1 that quite large time series models are needed to describe the sea water levels at the locations in question, also after long-term

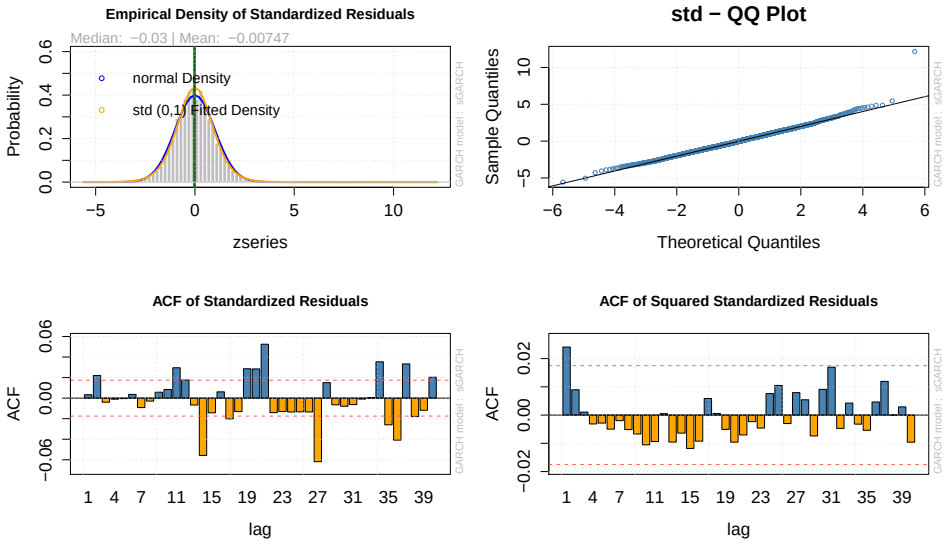

Figure 15: Diagnostic plots for the initial time series model fit to the Uddevalla data.

dependencies and heteroscedasticity are accounted for. One possible way to decrease the complexity of the model is to apply more pre-processing to the data, prior to modelling. There are clearly trends in the data that require more care. Trends and seasonality can either be modelled as deterministic or stochastic. In Section 5.3.1, the trend was treated as being stochastic, through the use of integrated ARMA models.

Based on the knowledge that the data in question is oceanographical, there should be a deterministic tidal seasonality in the data, as mentioned in Section 2.4. Estimating the tidal signal and subtracting it from the data prior to determining the time series model structure could lead to a more parsimonious model.

Classical tidal harmonic analysis is performed, as briefly described in Section 4.3.5. The `t_tide` package in MATLAB uses the Rayleigh resolution limit to decide which of 45 astronomical and 101 shallow-water tidal constituents to include in a least-squares fit. Also, nodal corrections are computed using the latitudes of the mobile gauges in question (see Section 2.1). For more details regarding the features of the package and the theory behind it, see Pawlowicz, Beardsley, and Lentz (2002).

Amplitudes and phases of the included constituents (as well as their respective errors) are computed, but not all are significant and should be used

to predict future tides. Pawlowicz, Beardsley, and Lentz (2002) suggest considering the tidal constituents with a signal to noise ratio (SNR) greater than two as significant. The SNR is computed as the squared ratio of the amplitude and the amplitude error for each constituent. Tables detailing which constituents that are significant for each of the three mobile gauge locations are presented in Tables 6, 7 and 8, together with their respective amplitude and phase estimates.

Tide	Freq.	Amp.	Amp_err.	Pha.	Pha_err	SNR
S_a	$1.141{\cdot}10^{-4}$	13.6937	7.273	279.61	29.31	3.5
O_1	0.0387	2.0616	0.568	298.29	15.14	13
EPS_2	0.0762	0.9706	0.606	257.67	34.21	2.6
$2N_2$	0.0775	1.0225	0.650	5.41	30.66	2.5
MU_2	0.0777	2.7050	0.599	291.64	11.52	20
N_2	0.0790	2.7192	0.509	84.94	12.00	29
NU_2	0.0792	1.3501	0.474	104.01	27.46	8.1
H_1	0.0804	1.0487	0.513	265.89	31.01	4.2
M_2	0.0805	13.8079	0.579	128.02	2.40	570
LDA_2	0.0818	0.8105	0.547	213.68	44.18	2.2
L_2	0.0820	1.8890	0.630	214.98	18.65	9
T_2	0.0832	0.8303	0.578	51.76	41.43	2.1
S_2	0.0833	3.1601	0.577	68.02	10.31	30
MO_3	0.1192	0.3949	0.230	339.66	33.47	2.9
MN_4	0.1595	0.4463	0.171	343.51	22.75	7.4
M_4	0.1610	1.7761	0.208	20.29	5.84	73
MS_4	0.1638	0.7974	0.189	93.66	14.80	18
$2MK_5$	0.2028	0.1248	0.051	118.14	29.74	5.9
$2MN_6$	0.2400	0.0986	0.034	24.17	21.26	8.6
M_6	0.2415	0.2139	0.037	56.92	9.42	34
$2MS_6$	0.2443	0.1835	0.038	135.85	10.56	24
M_8	0.3220	0.0751	0.023	182.61	18.92	11

Table 6: The significant (according to a cut-off rule at SNR < 2) tidal constituents for the Uddevalla data. The columns give the constituents names, frequencies (known), amplitude estimates and their errors, phase estimates and their errors, as well as their signal to noise ratio.

A trained oceanographer can most likely read more out of these tables, but three features stand out. The first is that there are a lot more significant

Tide	Freq.	Amp.	Amp_err.	Pha.	Pha_err	SNR
S_a	$1.141{\cdot}10^{-4}$	17.8811	5.359	253.31	18.91	11
O_1	0.0387	2.6841	0.960	306.74	20.55	7.8
$2N_2$	0.0775	0.8913	0.411	170.31	30.11	4.7
MU_2	0.0777	1.5163	0.354	28.27	13.50	18
N_2	0.0790	1.9728	0.390	171.73	12.11	26
NU_2	0.0792	0.6746	0.380	213.93	34.10	3.2
H_1	0.0804	0.5570	0.393	6.94	28.85	2
M_2	0.0805	7.4853	0.374	222.63	2.84	400
L_2	0.0820	0.4820	0.294	285.62	36.80	2.7
S_2	0.0833	2.0251	0.368	173.59	10.93	30
M_4	0.1610	0.3430	0.102	162.81	19.02	11
$2MN_6$	0.2400	0.1724	0.069	130.77	18.62	6.3
M_6	0.2415	0.3488	0.066	146.56	11.72	28
$2MS_6$	0.2444	0.3590	0.073	231.74	10.20	38
$2MK_6$	0.2446	0.1358	0.073	234.98	31.85	3.5
M_8	0.3220	0.0619	0.043	327.61	44.51	2.1

Table 7: The significant (according to a cut-off rule at SNR < 2) tidal constituents for the Ängelholm data. The columns give the constituents names, frequencies (known), amplitude estimates and their errors, phase estimates and their errors, as well as their signal to noise ratio.

constituents for the Uddevalla and Ängelholm data sets than for the Åhus data, as might be predicted considering the differences in their ACF plots in Figure 5. The second is that the annual solar constituent S_a is significant in the Uddevalla and Ängelholm analyses, but not the Åhus analysis. The third is that both the amplitudes and SNRs of the constituents are generally higher for the Uddevalla and Ängelholm analyses than for the Åhus analysis. E.g. the amplitude of the principal lunar semidiurnal constituent, M_2 is nearly 14 cm in Uddevalla while it is only 1 cm in Åhus, which is consistent with information given by SMHI at smhi.se (2013g).

This becomes very visible when tidal predictions based on the significant constituents are plotted. In Figure 16 four time series are plotted. The first two plots show tidal predictions over the time periods uses as modelling data for the time series models, for Uddevalla and Åhus. Comparing these two plots shows the impact of the annul solar constituent S_a on the tidal structure in Uddevalla clearly. Also, the heights of the tidal waves in Uddevalla and

Tide	Freq.	Amp.	Amp_err.	Pha.	Pha_err	SNR
O_1	0.0387	0.9595	0.605	279.39	38.00	2.5
K_1	0.0418	1.1040	0.605	269.15	38.77	3.3
$2N_2$	0.0775	0.2475	0.171	232.22	42.06	2.1
MU_2	0.0777	0.2649	0.166	152.82	38.92	2.6
M_2	0.0805	0.9953	0.188	54.14	8.70	28
S_2	0.0833	0.2677	0.182	57.26	33.10	2.2

Table 8: The significant (according to a cut-off rule at SNR < 2) tidal constituents for the Åhus data. The columns give the constituents names, frequencies (known), amplitude estimates and their errors, phase estimates and their errors, as well as their signal to noise ratio.

Åhus are noticeably different. The range of the Uddevalla tidal signal is slightly over 70 cm, while it is only 6.3 cm for the tidal signal in Åhus. This range difference can also be seen in the lower two plots, which show the first four days of each of these tidal time series. Here, another difference between the two tidal structures becomes apparent. Here, again, the findings of this tidal analysis are consistent with the information from SMHI at smhi.se (2013g), as the tide in Uddevalla is semidiurnal (Swe. halvdagligt). In Åhus, on the other hand, the tidal signal is a so called mixed semidiurnal tide, since there are two high tides of different heights per day.

5.3.3 Models for the post-tidal signal

With the aim of being able to diminish the complexity of the time series models designed in Section 5.3.1, the tidal signals calculated in the previous section are removed from the three data sets. These pre-processed data sets are used to build new time series models. Again, the Uddevalla series will be presented in the main text.

There are some differences to the model built to the pre-processed Uddevalla series, compared to the earlier model. One difference is that even though the tests used to study the presence of long-term memory in the data (the Elliott-Rothenberg-Stock test for unit roots and the Kwiatkowski-Phillips-Schmidt-Shin test for stationarity) still indicate that fractional integration could be helpful, neither the BIC nor visual inspections of the ACF/PACF of the model residuals indicate that a fractionally integrated model is preferable to a model with an integer valued d parameter. Both test and ACF-plots still show the presence of ARCH/GARCH effects. This

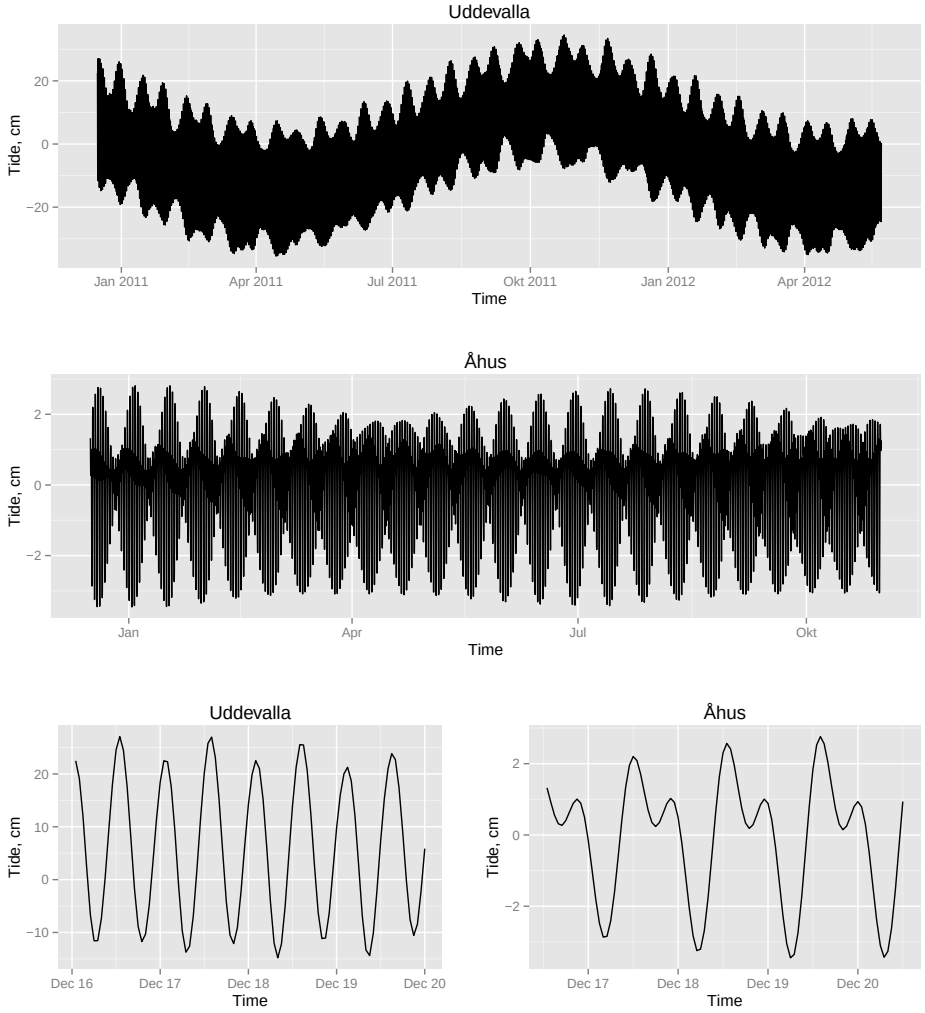

Figure 16: A comparison between the tides in Uddevalla and Åhus. The first two plots show tidal predictions over the time periods uses as modelling data for the time series models, for Uddevalla and Åhus. The lowest two show the first four days of each of these tidal time series.

indicates that a suitable model has an ARIMA-GARCH structure.

The final model for the pre-processed data is an ARIMA-GARCH(14,1,13)-(1,2) model. Once more, most of the ARMA parameters are set to zero. The included parameters are a_1, a_3, a_4, a_6, a_7, a_{10}, a_{12}, a_{14}, c_9, c_{12}, c_{13}. Clearly, this is a smaller model than the initial model previously presented, though still large. The largest simplification achieved is the removal of the

fractional integration. Figure 17 shows some diagnostics for the fit. Unfortunately, the ACF plot is very similar to Figure 15.

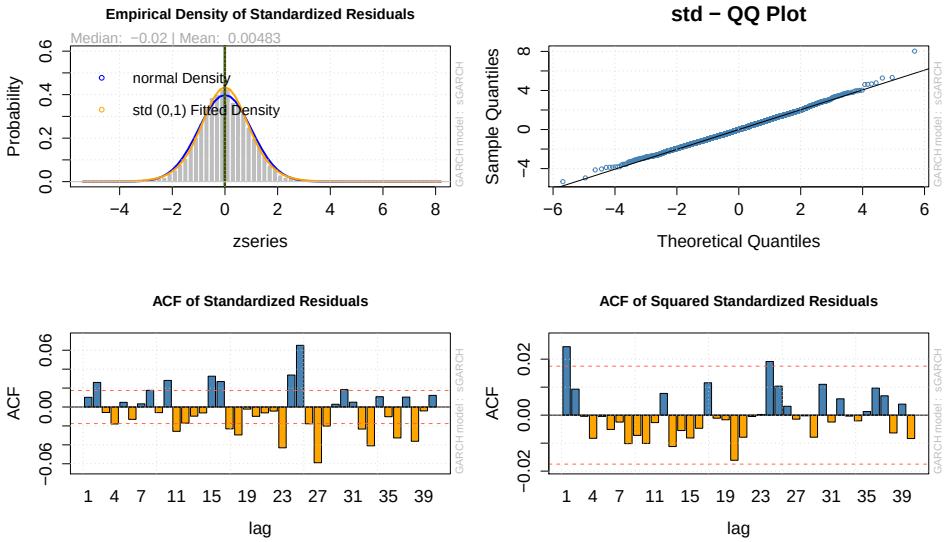

Figure 17: Diagnostic plots for the time series model fit to the pre-processed Uddevalla data.

To connect the time series models to the problems with temporal dependence in the data presented in the regression results section, time series models are also built for the data from the permanent gauges. Using the residuals from the Uddevalla and Smögen models the correlation between the two data sets can be compared, after the mean structure has been accounted for by the time series models.

The residuals have a linear correlation of 0.18. This can be compared to the correlation between the raw sea levels from Uddevalla and Smögen, 0.97. Clearly, a lot of the correlation between the data sets came from the mean structure. This is quite natural since the tidal signals are very similar for the two locations, given their geographical closeness. A linear regression model can be built using the residuals, instead of the raw data, removing the problems with temporal dependence in the data. Using BIC and stepwise F-tests, a multiple regression model is built:

$$\begin{aligned}\hat{y}_{\text{Udd,res}} =& \hat{\alpha} + \hat{\beta}_1 x_{\text{Smö,res}} + \hat{\beta}_3 x_{\text{ws}} + \hat{\beta}_4 x_{\text{S}} + \hat{\beta}_5 x_{\text{W}} \\ =& -0.16 + 0.14 x_{\text{Smö,res}} + 0.05 x_{\text{ws}} + 0.18 x_{\text{S}} - 0.18 x_{\text{W}}.\end{aligned}$$

The signs of the β parameters once more make sense, considering the geography of the locations, but the R^2 is only 0.038, indicting that quite a small amount of the variability in the residuals can be explained by the independent variables, regardless of them being significant in a regression model. Also, fewer parameters are included in this model than the previous regression models. Time series models and similar regression models can be found in Appendix D

6 Implementation

The analysis in this project has been done using the free statistical computing and graphics language `R` (R Core Team 2013). The basic `R` environment can be extend by installing user-written packages. Besides the packages that are distributed with the standard `R` download, several other packages have been used in this project. Some of the plots in the report have been produced using the graphical package `ggplot2` which is inspired by Leland Wilkinson's Grammar of Graphics (Carslaw and Chang 2013). The wind rose plot, Figure 8, was created using the `openair` package, which provides tools to analyse, interpret and understand meteorological data (Wickham and Ropkins 2013). The extreme value theory part of the project was made using the `evd` package (Stephenson and Ferro 2012). One alternative to the `evd` package is called `extRemes`. Both packages have their advantages, but the `evd` package was chosen for its ability to handle missing data.

In the time series analysis part of the project, the main package used is called `rugarch`. `rugarch` provides a large set of methods to fit, forecast and diagnose a huge variety of univariate ARFIMAX-GARCH models (Ghalanos 2013). The tests used for indicating the existence of long-term memory in the data (the Elliott-Rothenberg-Stock test for unit roots and the Kwiatkowski-Phillips-Schmidt-Shin test for stationarity) as well as the test for Arch/Garch effects (the McLeod-Li Test for conditional heteroscedascity) can be found in the `fUnitRoots`, `TSA` and `tseries` packages (Wuertz 2013; Chan and Ripley 2012; Trapletti and Hornik 2013).

One part of the time series analysis was done in MATLAB, namely the harmonic tidal analysis. This was done with the `t_tide` package which is based on the FORTRAN IOS Tidal package by Mike Foreman from 1977-78 (Pawlowicz, Beardsley, and Lentz 2002). There are a few MATLAB packages that compute harmonic analysis, the `t_tide` package was chosen for its simplicity and ability to handle missing values.

7 Summary and discussion

As was stated in Section 1, the Swedish Meteorological and Hydrological Institute (SMHI) placed mobile sea level gauges at places that were potentially oceanographically interesting. They wished to perform an extended statistical study of the results from these gauges. They also wished to relate the sea level data from the mobile gauges with the sea level data from nearby permanent stations in order to eventually be able to increase the reliability of the oceanographic forecasting and warning service. Also, they have an interest in combining sea level data with wind and atmospheric pressure data to better understand what factors affect the sea levels at the three locations of interest.

A statistical study of the data at hand was performed in Sections 3.1–3.2 as well as Section 5.1, where extreme value theory was used to study how large the differences between the three paired locations were likely to become. More specifically, five-year return levels were studied. In Table 3 these five-year return levels are presented. It can be seen that the return levels all roughly fall between 30 and 53 cm, indicating that the possible sea level differences between the paired locations can become quite large. It should be noted that the 95 % confidence intervals have varying sizes, depending on the quality of the GEV distribution fit, but some of them indicate that the five-year return levels could actually be as large as 70-80 cm. As is pointed out in Ch. 3 of Coles (2001), the use of GEV models for extrapolation is based on unverifiable assumptions, and thus return level confidence intervals should be regarded as lower bounds of the uncertainty, if model correctness could be accounted for. With this in mind, whether the sizes of the possible differences between the permanent and mobile sea level gauges are large enough to merit further attention is left to SMHI to judge. They are better equipped to conclude if these differences are larger than can be accepted, and thus if the mobile gauges are to be kept where they are, or moved.

One possible objection to the methodology of applying GEV models to the difference data is the possibility that the largest differences could be due to the tides at the paired locations being out-of-phase – that is that the sea level in e.g. Uddevalla could be at high tide (the peak of a sine movement) while the se level in Smögen is at low tide (the bottom of a sine movement). Thorough visual inspections of the points that correspond to the largest differences show that this is not the case. The tidal signals at the paired stations are similar enough not to cause large differences by being out-of-phase. Another possible objection is that it is not certain that the extremes

of e.g. Uddevalla – Smögen are really points where the Uddevalla sea level is higher than the Smögen sea level – it might just as well be points where the both sea levels are negative, and Smögen sea level is more extreme. While this is certainly possible, and probably true for some points, a thorough visual inspection of the most extreme (both negative and positive) points shows that this is not the case for the most extreme, and thus important, points.

A possible way to improve on the extreme value analysis could be to fit a Generalised Pareto Distribution, instead of the GEV distribution. The GPD has some positive aspects, when compared to GEV models, but since the GEV fits where generally good, a GPD fit was not attempted during this project.

Relating the sea levels from the mobile gauges with their permanent counterparts was done with linear regression in Section 5.2. It was largely successful, as shown by both diagnostic plots and quite correct prediction intervals, when the regressions were applied to validation data. The initial regression results were summarised in Table 4. It can be seen that the regression line for the Uddevalla – Smögen regression is the steepest of the four regressions and Åhus – Simrishamn is the flattest. Also, the Adj. R^2 was lowest for Uddevalla – Smögen and highest for Åhus-Simrishamn. This indicates that there is more behaviour in the Uddevalla data that cannot be explained by the sea levels in Smögen, than there is in the Åhus data that cannot be explained by the Simrishamn data. A slight quadratic structure in the relationship between the mobile stations and their permanent counterparts was also found. Additionally, the data from Kungsholmsfort was found to add some predictive power to the Åhus regression, but was discarded due to very high collinearity.

In order to examine the importance and influence of wind speeds, wind directions and atmospheric pressure on the behaviour of sea levels at the three locations, the linear regression was extended. The results were largely consistent with what was expected, considering the geography of the three locations. One interesting point is that the parameter representing the squared data from the permanent gauges are still significant after the meteorological data are added to the regression model. The largest breach of the regression assumptions was considered to be the temporal dependence in the residuals. This indicated that a time series model might be advantageous in describing the sea level behaviour. This is natural, since the data are structured as time series.

Thus, time series models were designed to describe the data in another manner, with the hope of gaining more information. Initially, ARMA models

where extended by adding GARCH structures to handle heteroscedasticity and fractional integration structure was added to better model long term memory in the time series. Unfortunately, very large ARFIMA-GARCH models were needed in order to get the residuals to be nearly white. This indicated the presence of a lot of structure in the data, which might be better handled with some sort of pre-processing. The oceanographical nature of the data indicated that suitable pre-processing might be the removal of tidal signals from the sea level data.

In studying the tidal signals, some structures stood out. The largest was the difference between the tidal signals on the west and east coasts. The tides in Åhus did not have a significant annual structure, in contrast to Uddevalla and Ängelholm. After removal of the tidal signals from the data, new time series models were built. One immediate effect of this pre-processing was that there no longer was as strong an indication for long-term memory as previously, possibly indicating that the long temporal structures handled by the fractional integration were in fact tidal patterns in the data.

One objective with building time series models was initially to be able to describe the mean structures of both the mobile and permanent gauge data sets and then use the residuals from their respective time series models to examine the relationships between the residuals from the paired stations. This could have been done using both regression theory again, or copula theory. Due to time constraints, the copula approach was not attempted. The copula approach could possibly have been more interesting, considering the lack of linear correlation between the residuals. This lack of correlation indicates that a lot of the similarities between the paired stations might be due to tidal structures and other men structures in the data captured by the time series models.

Another possible, and interesting, continuation of the time series modelling would be to combine the regression models with the time series models. This could be done through the use of exogenous regressors in so-called ARMAX models or by building a Generalised Least Squares model, which compensates for the correlation between the observations.

Deciding the future use of the mobile sea level gauge data is of course left to SMHI. This project has studied the differences between the paired data sets, as well as how they are related, though extreme value analysis and regression. The next step forward hinges on how the warning and forecasting work is done in practice at SMHI, when it comes to gauge data contra model data. One possible way to continue this study is to perform statistical post-processing on the model output from the models currently being used by the Oceanographic Warning & Forecasting Service, with the results presented

in this report as basis. Another possible next step could be to do model validation on the current models in use based on the findings in this project, but that hinges on the grid sizes used in the models, i.e. if they have fine enough grids to make a difference between e.g. Smögen and Uddevalla.

Appendices

A Mean Sea Level Equations

SMHI

2013-06-04

EKVATIONER FÖR
MEDELVATTENSTÅNDET I RH2000 **2013**
Beräknat medelvattenstånd i RH2000
cm

NR	STATION	EKVATION
2157	KALIX	Wyy=30.5-(0.73)*(yy-1986)
2055	FURUÖGRUND	Wyy=29.0-(0.82)*(yy-1986)
2056	RATAN	Wyy=30.3-(0.80)*(yy-1986)
2321	SKAGSUDDE	Wyy=27.1-(0.80)*(yy-1986)
2061	SPIKARNA	Wyy=24.6-(0.68)*(yy-1986)
2179	FORSMARK	Wyy=24.6-(0.64)*(yy-1986)
2069	STOCKHOLM	Wyy=21.5-(0.38)*(yy-1986)
2507	LANDSORT NORRA	Wyy=18.2-(0.29)*(yy-1986)
2076	MARVIKEN	Wyy=15.5-(0.18)*(yy-1986)
2080	VISBY	Wyy=13.4-(0.12)*(yy-1986)
2083	ÖLANDS NORRA UDDE	Wyy=15.8-(0.12)*(yy-1986)
2085	OSKARSHAMN	Wyy=15.0-(0.12)*(yy-1986)
2088	KUNGSHOLMSFORT	Wyy=13.5-(0.01)*(yy-1986)
2543	ÅHUS mobil	Wyy=12.9-(-0.08)*(yy-1986)
2320	SIMRISHAMN	Wyy=12.9-(-0.08)*(yy-1986)
30488	SKANÖR	Wyy=12.7-(-0.08)*(yy-1986)
2095	KLAGSHAMN	Wyy=11.3-(-0.06)*(yy-1986)
2099	BARSEBÄCK	Wyy=9.8-(-0.06)*(yy-1986)
2228	VIKEN	Wyy=4.7-(-0.10)*(yy-1986)
2542	ÄNGELHOLM mobil	Wyy=4.7-(-0.10)*(yy-1986)
2105	RINGHALS	Wyy=7.3-(0.10)*(yy-1986)
2109	GÖTEBORG-TORSHAMNEN	Wyy=8.1-(0.16)*(yy-1986)
2110	STENUNGSUND	Wyy=4.1-(0.17)*(yy-1986)
2541	UDDEVALLA mobil	Wyy=4.1-(0.17)*(yy-1986)
2111	SMÖGEN	Wyy=1.7-(0.18)*(yy-1986)
2130	KUNGSVIK	Wyy=1.8-(0.20)*(yy-1986)

där yy är årtalet

B GEV fits

B.1 Smögen – Uddevalla

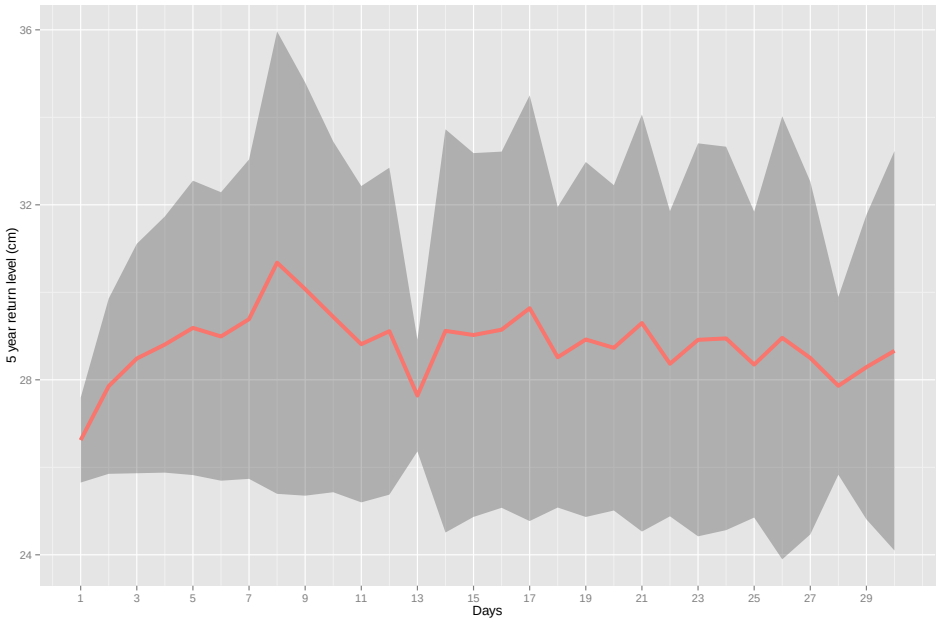

Figure 18: The five year return levels corresponding to GEV distributions fitted to block maxima of the difference between the Smögen measurements and the measurements from Uddevalla are shown by the red line. The block sizes are on the x axis, in multiples of 24 hours. A 95 % confidence band for the return levels is also shown, calculated by the delta method.

Choosing an appropriate block size for the Smögen - Uddevalla data is not as obvious as for the Uddevalla - Smögen case. Looking at Figure 18 it can be seen that the estimate of the five year return level is quite unstable for blocks smaller than two weeks. For block sizes larger than two weeks, the five year return level estimates seem to fall between 28 and 30 cm and the 95 % confidence intervals are more or less constant, covering the area between 24 and 34 centimetres. As can be seen in the ACF plot in Figure 19, the block maxima series (two-week blocks) appears white.

Looking at diagnostic plots for the GEV fit with two-week blocks, Figure 20, it can be seen that the fit is not as good as the previous Uddevalla - Smögen fit in Figure 11. One noticeable difference is that the negative value of the shape parameter ξ in this fit leads to a convex return level function,

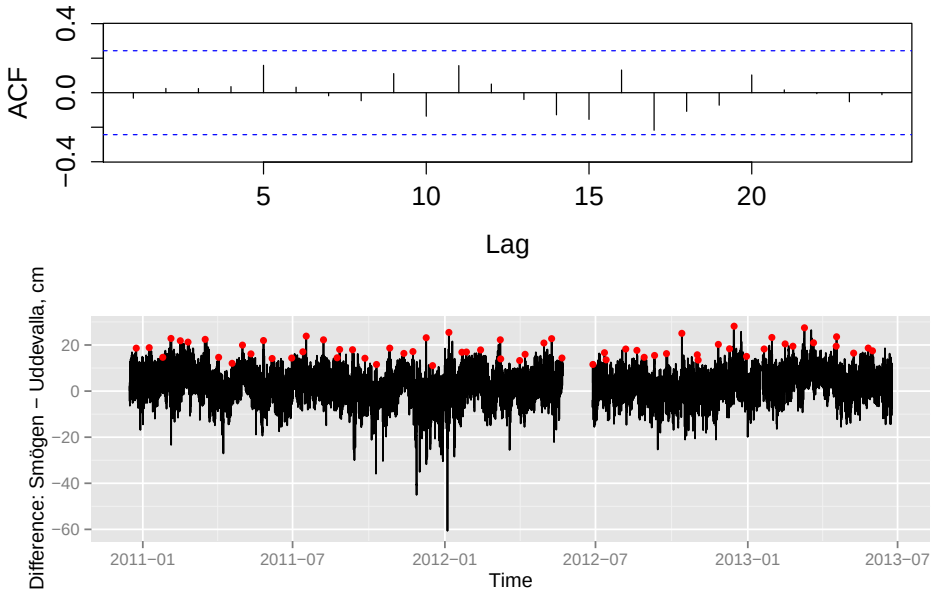

Figure 19: Bottom: a plot showing block maxima for the difference between the Smögen measurements and the Uddevalla measurements, with two-week blocks. Top: an autocorrelation plot for the same block maxima series.

which is more physically reasonable than a concave functions, since it implies a finite bound to the return levels. The GEV distribution parameters for this fit are location: $\mu = 16.61$ (0.50), scale: $\sigma = 3.55$ (0.36) and shape: $\xi = -0.15$ (0.10). The values in parentheses are standard errors. The length of the block maxima series used is 65.

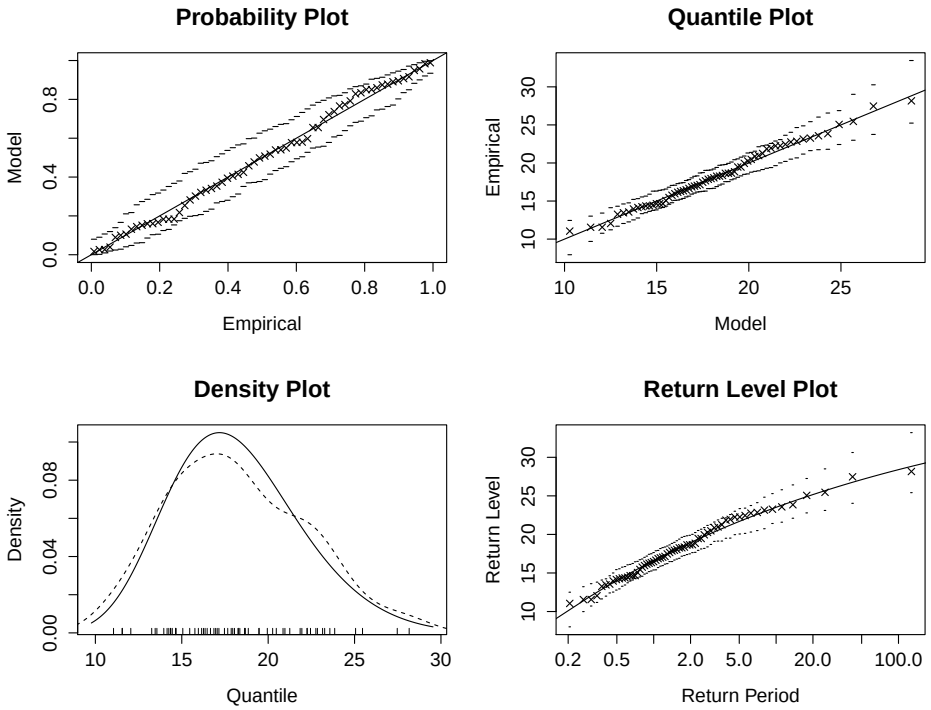

Figure 20: Diagnostic plots for the GEV fit to the Smögen - Uddevalla difference data. Left-to-right and top-to-bottom: 1) A probability plot. The data point follow a straight line reasonably well, considering the 95% confidence band. 2) A quantile-quantile plot. The plot further implies that the fitted model suits the data, since the model-based estimate of the quantile function fits the data well, considering the confidence bands. 3) The density of the fitted distribution is plotted (solid line), together with a non-parametric estimate (dashed line) and a rug plot of the data. The fit is not quite as good as the previous Uddevalla -Smögen fit. 4) The return level plot also shows that the fit is reasonably good, since the data points lie close to the return level line. The x-axis shows return levels (in multiples of 14 days) plotted on a logarithmic scale. The line is convex (and thus has a finite bound), which corresponds to the negative value of ξ in the GEV-fit.

B.2 Ängelholm - Viken and Viken - Ängelholm

A GEV distribution is fitted to the Ängelholm – Viken difference data, in the same manner as for the Uddevalla – Smögen case. The five year return level plot is shown in Figure 21. Using it together with ACF plots, a block size of one week is chosen. The block maxima autocorrelation is shown in Figure 22. 82 one-week block maxima are used for the fit. The GEV distribution parameters are location, $\mu = 4.54$ (0.46), scale, $\sigma = 4.34$ (0.34) and shape $\xi = 0.092$ (0.062). The shape parameter is positive, if not by much, indicating that the return level function has no finite bound. This should be considered before using this fit for very long return periods.

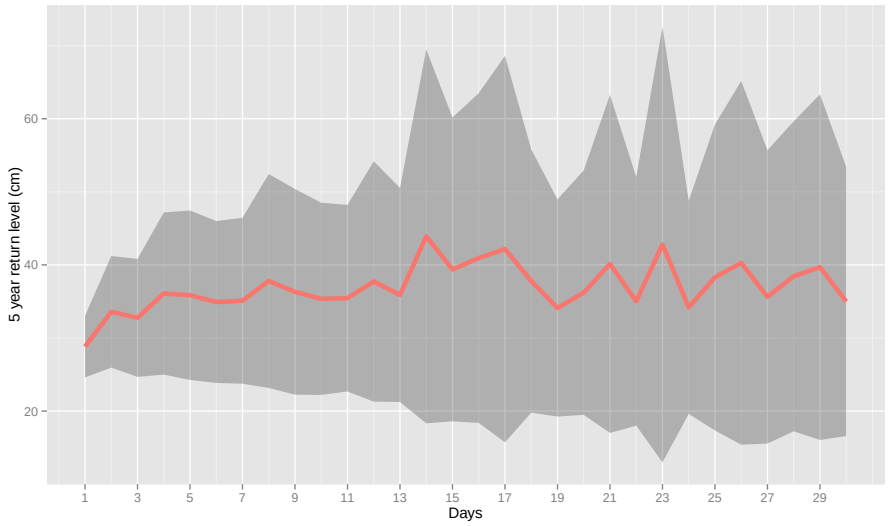

Figure 21: The five year return levels corresponding to GEV distributions fitted to block maxima of the difference between the Ängelholm measurements and the measurements from Viken are shown by the red line. The block sizes are on the x axis, in multiples of 24 hours. A 95 % confidence band for the return levels is also shown, calculated by the delta method.

All four diagnostic plots in Figure 23 show that the GEV fit is good. Especially the density plot is very promising.

A GEV distribution is fitted to the Viken – Ängelholm difference data (min of Ängelholm – Viken). The five year return level plot is shown in Figure 24. Using it together with ACF plots, a block size of 12 days is chosen. The block maxima autocorrelation is shown in Figure 25. 68 12-day block maxima are used for the fit. The GEV distribution parameters

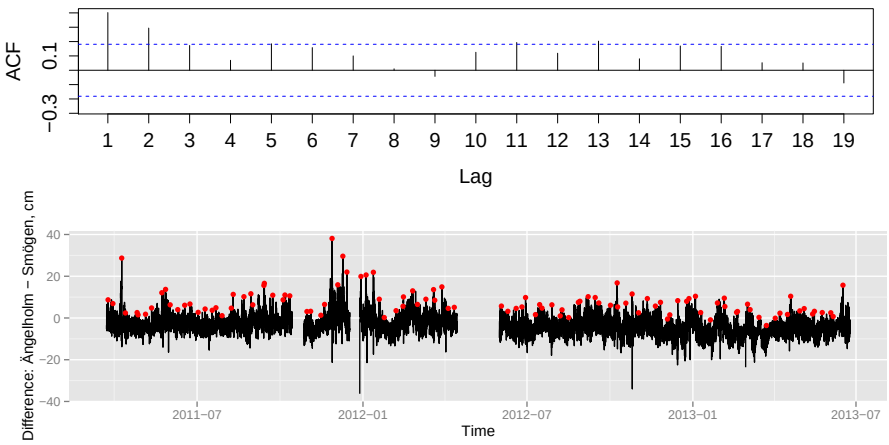

Figure 22: Bottom: a plot showing block maxima for the difference between the Ängelholm measurements and the Viken measurements, with seven-day blocks. Top: an autocorrelation plot for the same block maxima series.

are location, $\mu = 11.98$ (0.41), scale, $\sigma = 3.00$ (0.32) and shape $\xi = 0.16$ (0.09). The shape parameter is positive, again indicating that the return level function has no finite bound.

Again, four diagnostic plots are presented in Figure 26. They show that the GEV fit is good, though not quite as good as the one present in Figure 23.

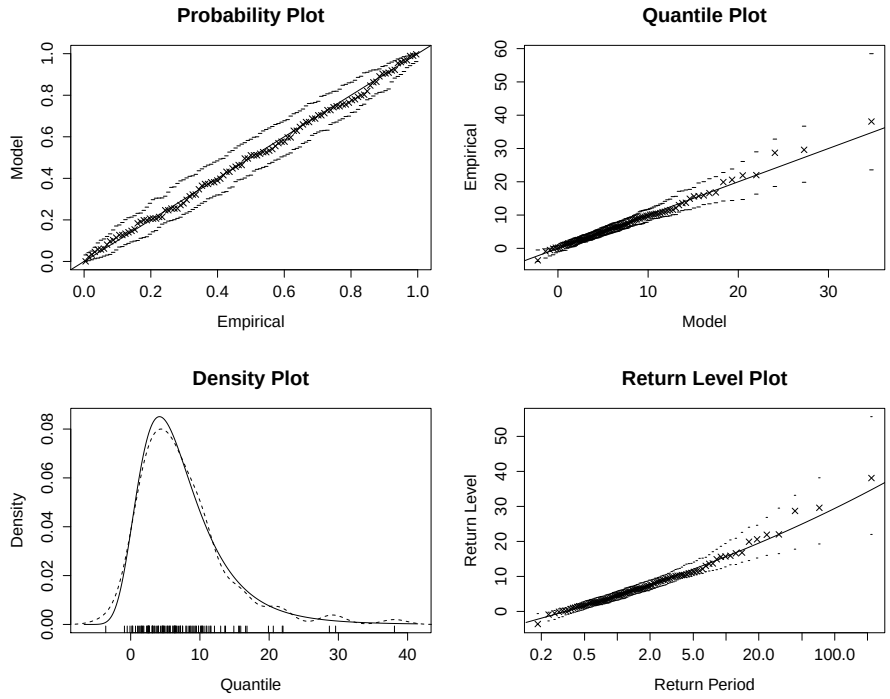

Figure 23: Diagnostic plots for the GEV fit to the Ängelholm - Viken difference data. Left-to-right and top-to-bottom: 1) A probability plot. The data point follow a straight line reasonably well, considering the 95% confidence band. 2) A quantile-quantile plot. The plot further implies that the fitted model suits the data, since the model-based estimate of the quantile function fits the data well, considering the confidence bands. 3) The density of the fitted distribution is plotted (solid line), together with a non-parametric estimate (dashed line) and a rug plot of the data. The fit is very good. 4) The return level plot also shows that the fit is good, since the data points lie close to the return level line. The x-axis shows return levels (in multiples of 7 days) plotted on a logarithmic scale. The line is concave (and thus has no finite bound), which corresponds to the positive value of ξ in the GEV-fit.

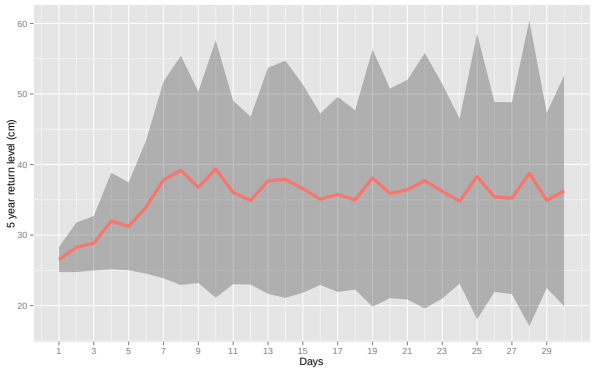

Figure 24: The five year return levels corresponding to GEV distributions fitted to block maxima of the difference between the Viken measurements and the measurements from Ängelholm are shown by the red line. The block sizes are on the x axis, in multiples of 24 hours. A 95 % confidence band for the return levels is also shown, calculated by the delta method.

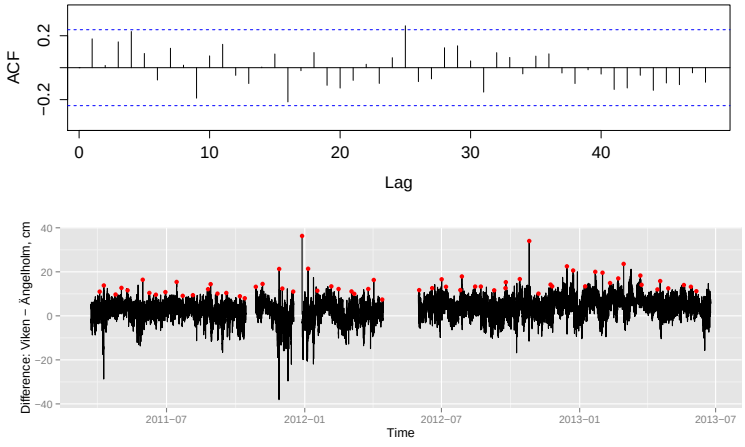

Figure 25: Bottom: a plot showing block maxima for the difference between the Viken measurements and the Ängelholm measurements, with 12-day blocks. Top: an autocorrelation plot for the same block maxima series.

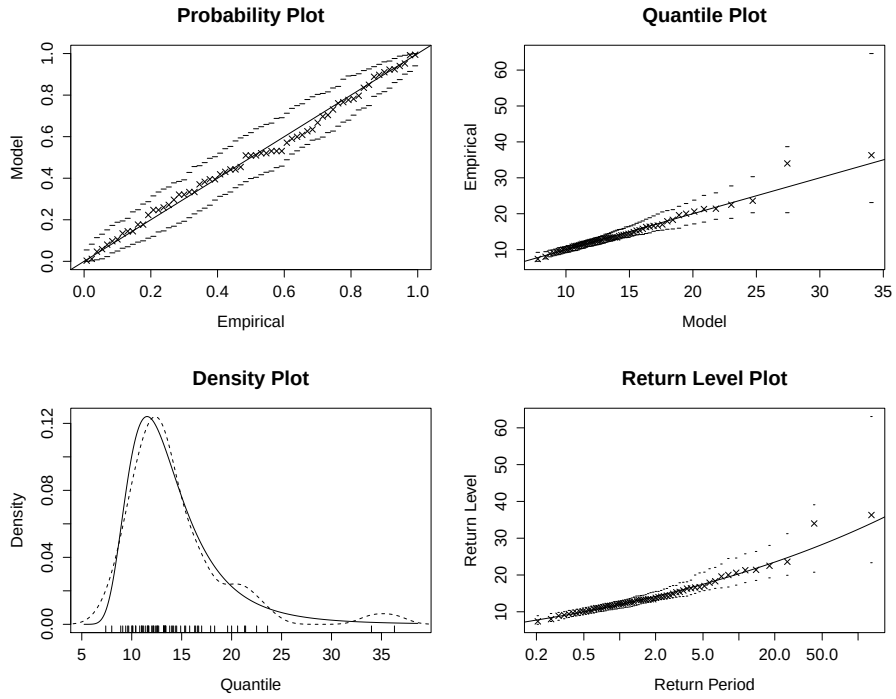

Figure 26: Diagnostic plots for the GEV fit to the Viken - Ängelholm difference data. Left-to-right and top-to-bottom: 1) A probability plot. The data point follow a straight line reasonably well, considering the 95% confidence band. 2) A quantile-quantile plot. The plot further implies that the fitted model suits the data, since the model-based estimate of the quantile function fits the data well, considering the confidence bands. 3) The density of the fitted distribution is plotted (solid line), together with a non-parametric estimate (dashed line) and a rug plot of the data. The fit is quite good, but has to low an upper tail to account for the largest block maximum. 4) The return level plot also shows that the fit is reasonably good, since the data points lie close to the return level line. The x-axis shows return levels (in multiples of 12 days) plotted on a logarithmic scale. The line is concave (and thus has no finite bound), which corresponds to the positive value of ξ in the GEV-fit.

B.3 Åhus - Simrishamn and Simrishamn - Åhus

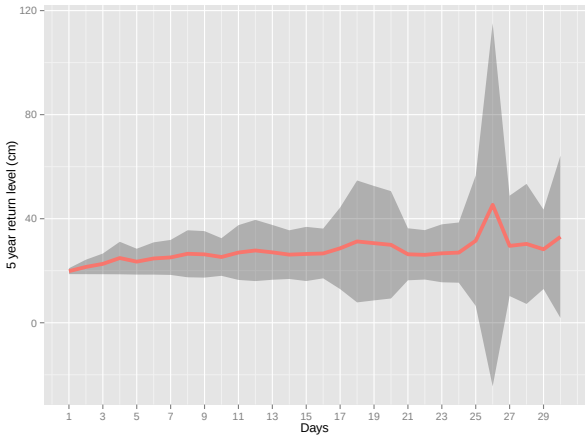

Figure 27: The five year return levels corresponding to GEV distributions fitted to block maxima of the difference between the Åhus measurements and the measurements from Simirishamn are shown by the red line. The block sizes are on the x axis, in multiples of 24 hours. A 95 % confidence band for the return levels is also shown, calculated by the delta method.

As previously, a GEV distribution is fitted to the Åhus – Simrishamn difference data. The five year return level plot is shown in Figure 27. Using it together with ACF plots in the same manner as before, a block size of one week is chosen. The block maxima autocorrelation is shown in Figure 28. The five year return level plot is quite different, visually, from the previous return level plots. It appears to not be especially biased, ever for small block sizes, and it is quite stable up to block sizes of nearly a month.

75 one-week block maxima are used for the fit. The GEV distribution parameters are location, $\mu = 7.84$ (0.36), scale, $\sigma = 2.83$ (0.26) and shape $\xi = 0.03$ (0.07). The shape parameter is positive, again indicating that the return level function has no finite bound. It should be observed that the standard error of the ξ parameter indicates that this is very uncertain. The return level function shown in Figure 29 is very nearly linear.

Once more, a GEV distribution is fitted to the minima of the difference data, in this case Simrishamn – Åhus. The five year return level plot is shown in Figure 30. Using it together with ACF plots in the same manner as before, a block size of one week is again chosen. The block maxima autocorrelation is shown in Figure 31. The five year return level plot is once morequite different, visually, from the previous return level plots. This time it is a lot

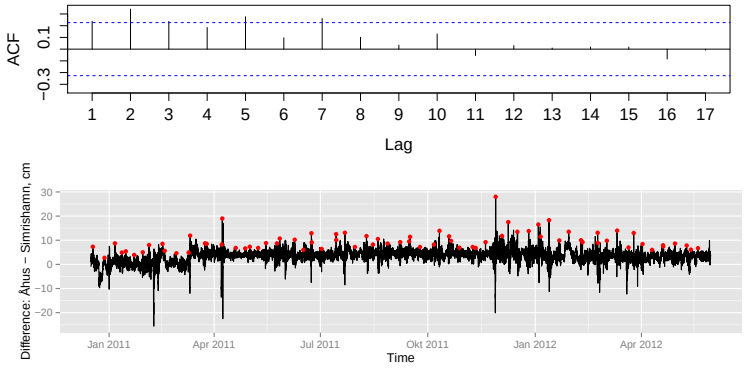

Figure 28: Bottom: a plot showing block maxima for the difference between the Åhus measurements and the Simrishamn measurements, with one-week blocks. Top: an autocorrelation plot for the same block maxima series.

more unstable, and appears to not centre around a small interval.

75 one-week block maxima are used for the fit. The GEV distribution parameters are location, $\mu = 1.16$ (0.38), scale, $\sigma = 2.81$ (0.33) and shape $\xi = 0.33$ (0.12). The shape parameter is positive, again indicating that the return level function has no finite bound, this time quite clearly.

The diagnostic plots in Figure 32 show that this is (possibly as expected, considering Figure 30) the least good fit of all. Especially the density plot is clearly not on par with those of the previous fits.

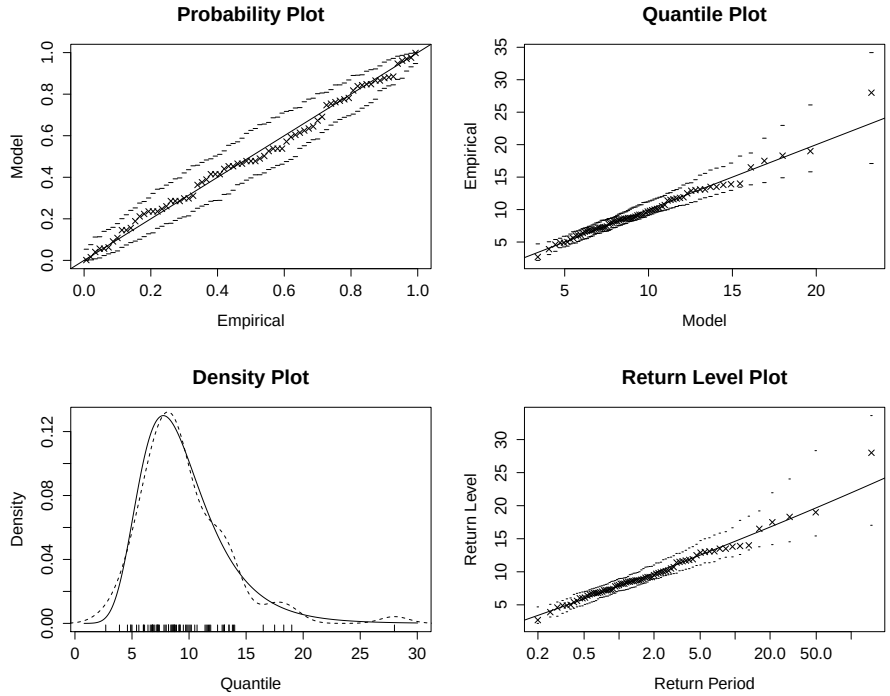

Figure 29: Diagnostic plots for the GEV fit to the Åhus - Simrishamn difference data. Left-to-right and top-to-bottom: 1) A probability plot. The data point follow a straight line reasonably well, considering the 95% confidence band. 2) A quantile-quantile plot. The plot further implies that the fitted model suits the data, since the model-based estimate of the quantile function fits the data well, considering the confidence bands. 3) The density of the fitted distribution is plotted (solid line), together with a non-parametric estimate (dashed line) and a rug plot of the data. The fit is very good. 4) The return level plot also shows that the fit is reasonably good, since the data points lie close to the return level line. The x-axis shows return levels (in multiples of one week) plotted on a logarithmic scale. The line is nearly straight, and only slightly concave compared to the other fits, which corresponds to the lower, but still positive, value of ξ in the GEV-fit.

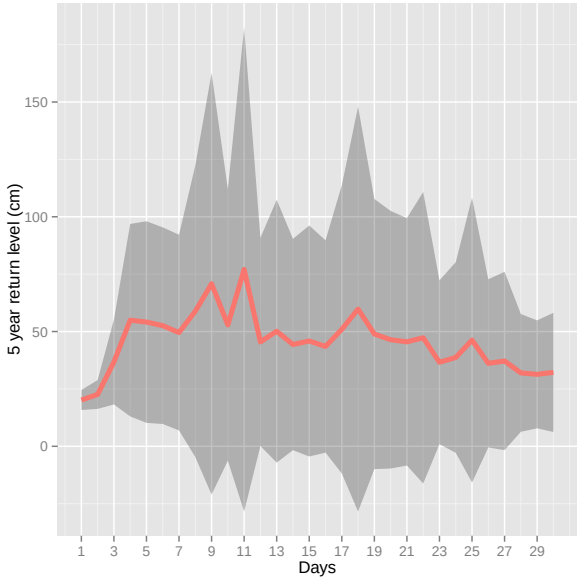

Figure 30: The five year return levels corresponding to GEV distributions fitted to block maxima of the difference between the Simrishamn measurements and the measurements from Åhus are shown by the red line. The block sizes are on the x axis, in multiples of 24 hours. A 95 % confidence band for the return levels is also shown, calculated by the delta method.

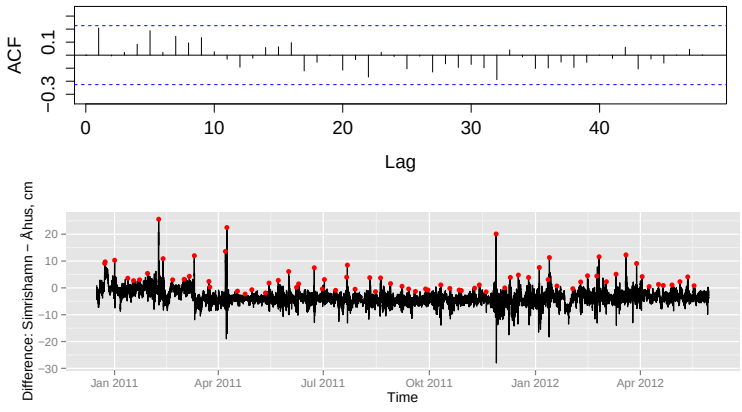

Figure 31: Bottom: a plot showing block maxima for the difference between the Simrishamn measurements and the Åhus measurements, with one-week blocks. Top: an autocorrelation plot for the same block maxima series.

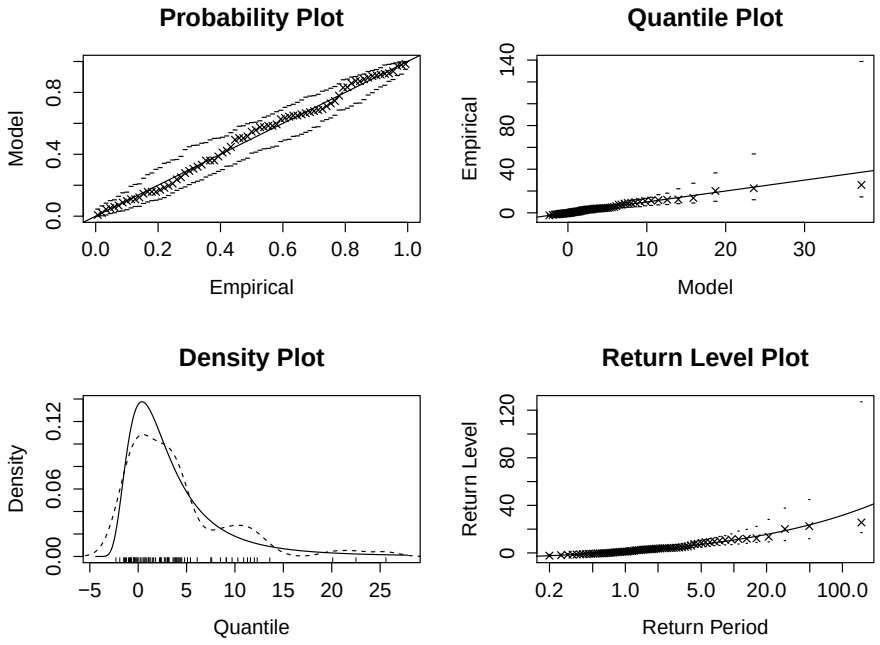

Figure 32: Diagnostic plots for the GEV fit to the Simrishamn - Åhus difference data. Left-to-right and top-to-bottom: 1) A probability plot. The data point follow a straight line reasonably well, considering the 95% confidence band. 2) A quantile-quantile plot. The plot further implies that the fitted model suits the data, since the model-based estimate of the quantile function fits the data well, considering the confidence bands. 3) The density of the fitted distribution is plotted (solid line), together with a non-parametric estimate (dashed line) and a rug plot of the data. The fit is not nearly as good as for the Åhus - Simrishamn fit in Figure 29. 4) The return level plot also shows that the fit is reasonably good, since the data points lie close to the return level line. The x-axis shows return levels (in multiples of one week) plotted on a logarithmic scale.

C Linear regression models

C.1 Regression diagnostics

Here, regression diagnostics for Ängelholm, C.1.1, and Åhus, C.1.2, are presented. In C.2, parameter estimates for the three best multiple regression models are given.

C.1.1 Ängelholm

As for the Uddevalla diagnostics presented in the main text, there is a quadratic trend visible in the plot of the regression residuals e_i versus the fitted values $\hat{Y}_i$ in Figure 34. The Cook's distance values for the residuals of the initial regression residuals for Ängelholm are not overly large. In contrast to the diagnostics for Uddevalla, the same data points are not both residual outliers and influential points (according to Cook's distance) for the Ängelholm regression.

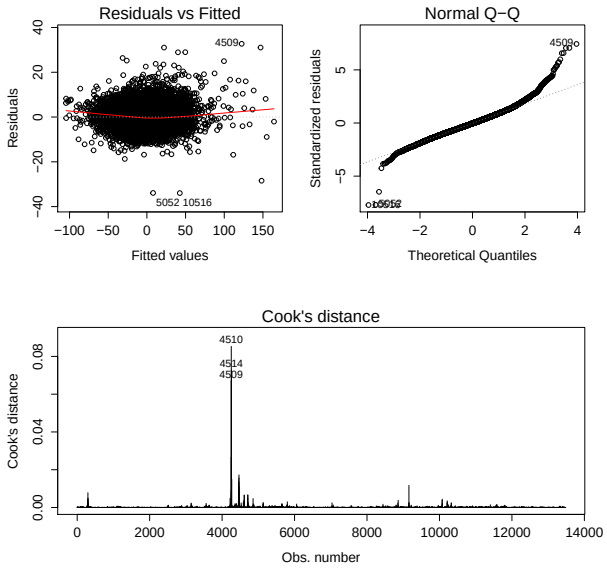

Figure 33: Top-to-bottom and left-to-right: a plot of regression residuals vs. the fitted values, a normal quantile-quantile plot, and a plot showing Cook's distance.

In the same manner as for the Uddevalla regression, the points indicated by the diagnostic plots are studied in more detail. From visual inspection of them, the same conclusions are drawn as for the indicated data points in the

Uddevalla regression. They are not obviously faulty measurements, as the points indicated in Figure 3.1.1 are, and are kept in the data set, hopefully to be better described in extended models.

C.1.2 Åhus

The slight curved structure of the regression residuals is the same for this fit as for the previous initial regression models. Also as previously, the normal quantile-quantile plot is not perfect, but this does not appear to cause the confidence and prediction intervals to be too biased, as is discussed in Section 5.2.1. As previously, closer inspections of the possibly troublesome points in the data set indicates that they should be left as they are.

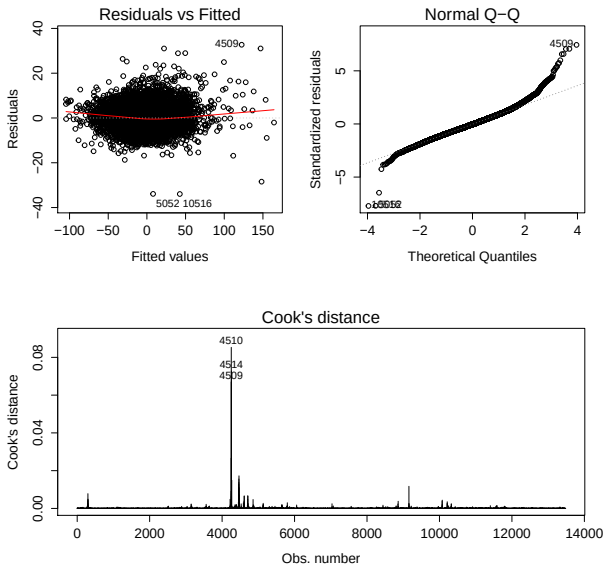

Figure 34: Top-to-bottom and left-to-right: a plot of regression residuals vs. the fitted values, a normal quantile-quantile plot, and a plot showing Cook's distance.

C.2 Multiple regression parameter estimates

	U–S	Ä–V	Å–S
$\hat{\alpha}$	-2.935 (0.1312)	32.89 (4.226)	22.02 (2.659)
$\hat{\beta}_{\text{Perm.}}$	1.093 (0.0027)	1.021 ($2.107{\cdot}10^{-3}$)	1.011 ($1.518{\cdot}10^{-3}$)
$\hat{\beta}_{\text{Perm.}^2}$	$-3.035{\cdot}10^{-4}$	$-3.322{\cdot}10^{-4}$	$-3.412{\cdot}10^{-4}$
	($6.069{\cdot}10^{-5}$)	($3.852{\cdot}10^{-5}$)	($3.661{\cdot}10^{-5}$)
$\hat{\beta}_{\text{Ws.}}$	0.3061 (0.0327)	0.4522 (0.0234)	-0.0751 (0.0126)
$\hat{\beta}_{\text{N}}$		-1.434 (0.1705)	-4.378 (0.1168)
$\hat{\beta}_{\text{NE}}$	-2.777 (0.1542)	-3.025 (0.1589)	
$\hat{\beta}_{\text{E}}$	-2.896 (0.1851)	-2.293 (0.1691)	1.681 (0.0945)
$\hat{\beta}_{\text{SE}}$	-2.852 (0.2235)	-1.102 (0.1565)	1.657 (0.0939)
$\hat{\beta}_{\text{S}}$	-1.224 (0.1449)	-0.6652 (0.1353)	1.925 (0.0896)
$\hat{\beta}_{\text{SW}}$		1.403 (0.1395)	1.441 (0.0731)
$\hat{\beta}_{\text{W}}$	1.026 (0.1725)	1.173 (0.1418)	
$\hat{\beta}_{\text{NW}}$			
$\hat{\beta}_{\text{Press.}}$		-0.0368	-0.0185
		($4.139{\cdot}10^{-3}$)	($2.605{\cdot}10^{-3}$)

Table 9: Parameter estimates for the three regression models. The standard error of the estimates are given in brackets.

D Time series models for Ängelholm and Åhus

Time series models are built for the data from Ängelholm and Åhus in the same way as for Uddevalla. After the tidal signals are removed, the model found for Ängelholm is an ARMA(3,0,25)-GARCH(1,1) model where the chosen ARMA parameters are: μ, a_1, a_3, c_3, c_4, c_9, c_{12}, c_{13}, c_{24}, c_{25}. Observe that an an ARMA model was preferred by both BIC and visual inspections of ACF/PACF plots to an ARIMA or ARFIMA model. Diagnostic plots for this fit are very similar to the plots for the Uddevalla fit, shown in Figure 17. Once again, a Student's-t distribution fits the residuals better than a Normal distribution. The residuals from this model have a linear correlation of 0.38 with the residuals from a time series model for Smögen. The best fitting linear regression model for the residuals has an R^2_{Adj} of 0.15 and is

$$\begin{aligned}\hat{y}_{\text{Äng,res}} =& \hat{\alpha} + \hat{\beta}_1 x_{\text{Vik,res}} + \hat{\beta}_2 x_{\text{E}} + \hat{\beta}_3 x_{\text{Pr.}} \\ =& 16.39 + 0.35 x_{\text{Vik,res}} - 0.25 x_{\text{E}} - 0.02 x_{\text{Pr.}}.\end{aligned}$$

For the Åhus data, the chosen model is slightly larger. It is an ARIMA(1,1,20)-GARCH(1,1) model with these chosen ARMA parameters: μ, a_1, a_2, a_3, a_4, a_5, c_7, c_9, c_{10}, c_{11}, c_{14}, c_{16}, c_{18}, c_{19}, c_{20}. The residuals from this model have a linear correlation of 0.45 with the residuals from a time series model for Smögen. The best fitting linear regression model for the residuals has an R^2_{Adj} of 0.20 and is simpler than the previous models:

$$\begin{aligned}\hat{y}_{\text{Åh,res}} =& \hat{\alpha} + \hat{\beta}_1 x_{\text{Shamn,res}} \\ =& 0.004 + 0.51 x_{\text{Shamn,res}}.\end{aligned}$$

There appears to still be some structure left in the data for Åhus and Simrishamn, even after the tidal signals have been removed and time series models have been applied.

References

Akaike, Hirotugu (1974). *A new look at statistical model identification*. Vol. 19. 6, pp. 716–723.

Baum, Christopher F. (2013). *ARIMA and ARFIMA models*. EC 823: Applied Econometrics Boston College.

Baum, Christopher F. and Vince Wiggins (2000). *sts16: Tests for long memory in a time series*. Stata Technical Bulletin 57: 39-44.

Blom, Gunnar et al. (2005). *Sannolikhetsteori och statistikteori med tillämpningar*. Swedish. 5:5. Studentlitteratur AB, Lund. ISBN: 978-91-44-02442-4.

Bollerslev, Tim (1985). *Generalized autoregressive conditional heteroscedasticity*. Memo / Økonomisk institut, Aarhus universitet: 1985:9. Aarhus.

Carslaw, David and Winston Chang (2013). *ggplot2: An implementation of the Grammar of Graphics*. R package version 0.9.3.1.

Chan, Kung-Sik and Brian Ripley (2012). *TSA: Time Series Analysis*. R package version 1.01.

Coles, Stuart (2001). *An introduction to statistical modeling of extreme values*. Springer series in statistics. London : Springer, cop. 2001. ISBN: 1852334592.

Cook, R. Dennis (1977). "Detection of Influential Observation in Linear Regression." In: *Technometrics* 19.1, p. 15.

Cryer, Jonathan and Kung-Sik Chan (2008). *Time Series Analysis : With Applications in R*. 2nd ed. Springer Texts in statistics. Springer.

Doodson, Arthur Thomas (1921). "The Harmonic Development of the Tide-Generating Potential". In: *Proceedings of the Royal Society of London. Series A, Containing Papers of Mathematical of Physical Character* 100.704, pp. 305–329.

Engle, Robert F. (1982). "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation". In: *Econometrica* 50.4, pp. 987–1007.

Faraway, Julian (2004). *Linear Models with R*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

Ghalanos, Alexios (2013). *rugarch: Univariate GARCH models*. R package version 1.2-7.

Jakobsson, Andreas (2013). *An Introduction to Time Series Modeling*. Swedish. 1:1. Studentlund AB, Lund. ISBN: 978-91-44-08374-2.

Jönsson, Anette and Robert Olsson (2013). *Instruktioner för vakthavande oceanograf*. Swedish. Dnr.:2013/643/2.10.1.

Leadbetter, M.R., G. Lindgren, and H. Rootzén (1983). *Extremes and Related Properties of Random Sequences and Series*. Springer Verlag, New York.

McLeod, A. I. and W. K. Li (1983). "Diagnostic checking ARMA time series models using squared residual autocorrelations". In: *Journal of Time Series Analysis* 4.4, p. 269.

Pawlowicz, Rich, Bob Beardsley, and Steve Lentz (2002). "Classical tidal harmonic alanysis including error estimates in MATLAB using T_TIDE". In: *Computers & Geosciences* 28.8, pp. 929–937.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Rawlings, John O., Sastry G. Pantula, and David A. Dickey (2001). *Applied regression analysis: a research tool*. 2nd ed. Springer texts in statistics. Springer Verlag, New York.

Schwartz, Gideon (1978). *Estimating the Dimension of a Model*. Vol. 6. 2, pp. 461–464.

smhi.se (Sept. 2013a). *Årets medelvattenstånd*. Swedish. URL: http://www.smhi.se/kunskapsbanken/oceanografi/arets-medelvattenstand-1.10047.

— (Sept. 2013b). *Havsvattenstånd*. Swedish. URL: http://www.smhi.se/kunskapsbanken/oceanografi/havsvattenstand-1.3090.

— (Sept. 2013c). *Havsvattenstånd i RH2000*. Swedish. URL: http://www.smhi.se/kunskapsbanken/oceanografi/havsvattenstand-i-rh2000-1.30859.

— (Sept. 2013d). *Havsvattenståndets årstidsvariationer*. Swedish. URL: http://www.smhi.se/kunskapsbanken/oceanografi/havsvattenstandets-arstidsvariationer-1.25961.

— (Sept. 2013e). *Lufttryck och havsvattenstånd*. Swedish. URL: http://www.smhi.se/kunskapsbanken/oceanografi/lufttryck-och-havsvattenstand-1.3096.

— (Sept. 2013f). *Smhi | Klimatdata | Oceanografi | Havsvattenstånd*. Swedish. URL: http://www.smhi.se/klimatdata/oceanografi/Havsvattenstand.

— (Oct. 2013g). *Tidvatten*. Swedish. URL: http://www.smhi.se/kunskapsbanken/oceanografi/tidvatten-1.321.

Stephenson, Alec and Chris Ferro (2012). *evd: Functions for extreme value distributions*. R package version 2.3-0.

Strömberg, Patrik (2012). *Utvärdering mobila vattenståndsmätare*. Swedish. DM#134215.

Tiku, Moti Lal (1971). "Power Function of the F-test under Non-Normal Situations". In: *Journal of the American Statistical Association* 66.336, pp. 913–916.

Trapletti, Adrian and Kurt Hornik (2013). *tseries: Time Series Analysis and Computational Finance*. R package version 0.10-32.

Vasilev, Oleg Fedorovich (2007). *Extreme Hydrological Events: New Concepts for Security: New Concepts for Security*. Nato Science Series: IV: Earth and environmental sciences. Physica-Verlag.

Wickham, Hadley and Karl Ropkins (2013). *openair: Tools for the analysis of air pollution data*. R package version 0.8-5.

Wuertz, Diethelm (2013). *fUnitRoots: Trends and Unit Roots*. R package version 3010.78.

Master's Theses in Mathematical Sciences 2014:E12
ISSN 1404-6342

LUTFMS-3242-2014

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lth.se/