Extreme value modeling of wind effect on dune erosion on the Coast of Ängelholm

Lionel Arpin-Pont

January 2020

Abstract

The movement of the coast line, due to erosion in one direction or aggregation in the other is a natural process as waves, wind as well as the geological nature of the coast itself are affecting it. Sand dunes are the main protection coasts have at their disposal against floods during storm surges or from more passive but long range rainfalls.

As many of the coastal areas have a dense population and it also shows a great biodiversity, the dune erosion is a phenomenon worth investigating since it is the destruction of such protection which is vital for everything living close-by. The cost of a flood can be measured by human loss, landscape damage or construction loss. Most of the buildings are not suited for floods.

Studying the dune erosion by itself might not be enough to provide good advice in case of a surge since the erosion is an effect of the surge as much as it then increases the following risks of flooding. In order to be well prepared and build efficient methods against such flooding, it is necessary to better understand the dune erosion and the surrounding phenomenon, such as the sea-level rise, the wave runups or the wind speed.

The data taken for this study comes mostly from the SMHI (Swedish Meteorological and Hyrological Institute), taken on or nearby the shore of Ängelholm in Skåne, south-west of Sweden.

Keywords : Erosion, Wind speed, Extreme value theory, Block maxima, Peaks over threshold, Copula, Husler-Reiss, Dependence function.

Acknowledgement

I would like to thank my supervisor, Nader Tajvidi in Lund University for his time, patience, help and advices throughout my master thesis and for giving me this subject about which I had never had been thinking of before, but which has interested me this whole semester.

I would also like to thank some of my teachers from Aix-Marseille University in France for the mathematical knowledge they passed down and the passion they inspired during my bachelor ; and my teachers from Lund University for the interest they gave me about statistics.

Finally, I would like to thank my family and friends for their support and help during my studies, and more importantly, during this thesis. A special thanks to Thomas Gourdel and Ellen Barrett for proofreading my report.

Last but not least, I want to thank my father who was always by my side and who introduced me to mathematics when I was younger with bed-time stories about strange attractors and cloud's trajectories models.

Table des matières

1 Introduction

In this master thesis report, the effect of wind on dune erosion will be dealt with. The data come from different stations around the beach of Ängelholm, Sweden.

Figure 1 – Picture of the coast of Ängelholm from Sanna Ny, on badkartan.se website.

A previous study was done on the effect of wave runups and sea level on the sand dunes of the same region by C. Hallin (2019), "Long-term beach and dune evolution, Development and application of the CS-model", see [**3**]. Using 59 different points along the shore over 40 years (between 1976 and 2015, included). The analysis will start there to study the effect of wind on dune erosion, to see if and how it amplifies the process. The analysis will be done using extreme value theory, first with univariate models to get a good grasp at how those phenomena behave and then with multivariate ones to see how they work together. The theory will be explained in section 2, entitled "Theoretical Background".

The goal of this thesis is to provide a better understanding of the influence and effect of the wind on dune erosion on the coast of Ängelholm in Sweden, using extreme value theory. It will be done by creating a bivariate extreme value model and find quantile for

return levels, i.e, the worst case scenario, that is the maximum erosion on the sand dunes in the future.
This will be done by finding the most suited data sets from stations surrounding the area and using extreme value theory to model the erosion using wind speeds and directions. In the end, using the results of the analysis of the different data sets, one will be able to see if there is a risk of flood and damage on the land surrounding Ängelholm's beach.

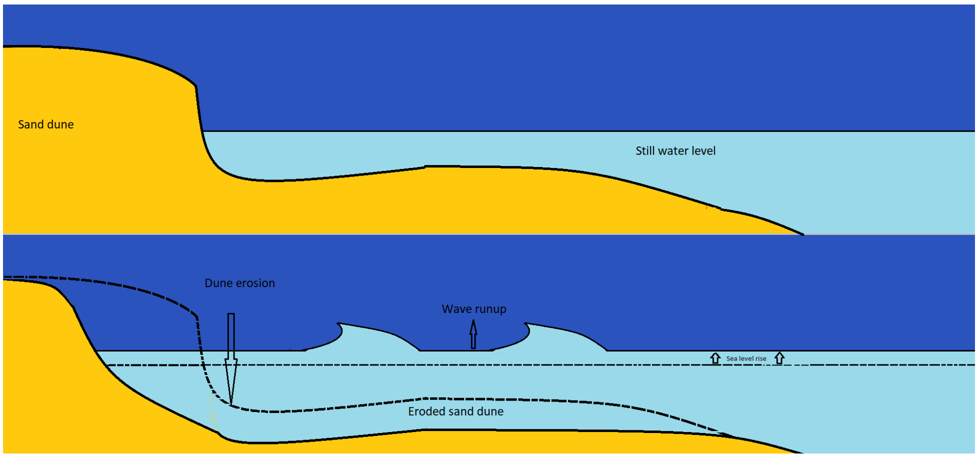

Figure 2 – Diagram of the erosion process with the dune erosion, sea-level rise and the wave run-up.

We start with an introduction to a theoretical part about stochastic processes, both univariate and multivariate extreme values models will be introduced as well as some important theory about copulas and goodness of fit methods. Then, dealing with the data itself : data can, and most of time is, difficult to use as it is. Between missing values or uncertain measurements, choices need to be made. There will be attempts to solve these problems. Afterwards, proceeding with the extreme value analysis of the different data sets, from sea-level, maximum wave runups and dune erosion to wind speeds and directions. At first, univariate models will be sought for as one wants to better understand the phenomena by themselves in order to predict flood risks. And later on, with the wind components, a bivariate model will be created for erosion.

During the analysis, the sets used will be the ones which match the best with the requirements regarding number of data, period matching, quality and so on, trying to solve the data-problem in our attempts.

2 Theoretical Background

Starting with stationary stochastic processes which are of the most important processes in statistical analysis. A random process is a sequence of random variables $X_1, X_2, ...$ They can either be dependent or not, identically distributed or not. A random process which has an homogeneous dependence in time is called a stationary process. Such processes are broadly studied and are defined below.
Definition : A random process $X_1, X_2, ...$ is stationary if, given integers $i_1, ...i_k$ and any integer m, the joint distribution of $\{X_{i_1}, ...X_{i_k}\}$ and $\{X_{i_1+m}, ..., X_{i_1+m+k}\}$ are identical for any choice of m.
In other terms, it implies that the mean and the variance of the process are constant over time and that the covariance function only depends on the shift and not in time, i.e the correlation function, defined as

$$\rho(a, b, t) = corr(X(t + a), X(t + b))$$

and must fulfil the following property

$$\rho(a, b, t) = \rho(\tau), \text{ where } \tau = b - a, \forall a, b \in \mathbb{N}, \forall t \in \mathbb{R}^+ .$$

2.1 Extreme Value Theory

To begin with, a short introduction to extreme value theory is presented, starting with the univariate models.

Extreme value theory can be used to model phenomena for which the number of occurrences is low, as in for example storm surges, floods, financial krach or else, etc. It doesn't require to know the distribution function of the occurring phenomena, but only that the distribution asymptotically converges towards an extreme value distribution. It is used to forecast events to come that are very unlikely but which can happen over a long period of time, and to prepare for its worst case scenario.

In this paper, extreme value theory will be used for coastal protection. Floods, when occurring, can damage coastal cities and landscapes as much as its population. For insurance and safety reasons, this is a topic of matter now more than in the past as with climate change, sea-levels rise and floods are even more impacting.

Extreme value analysis is then used to calculate return levels, i.e, levels of a data that can be reach in the worst events (as a storm surge or a flood). As there is not much data on those events, one can use asymptotic arguments to work with extreme values.

In one hand there is the block-maxima method. Let a random process $X_1, X_2, ..., X_n$ have

M_n as its maximum value. Knowing how the $\{X_i\}$'s behave would also give M_n's behavior, but as they are both unknown (for lack of data), one can't know M_n's exact distribution.

Hence the following approach

$$\mathrm{P}(\mathrm{X}_1 \leq x, X_2 \leq x, ...X_n \leq x) = \Pi_{i=1}^{n} P(X_i \leq x) = F^n(x).$$

F being unknown and G, the limiting distribution of F, is degenerate in $x^F = supx \in \mathbb{R} : F(x) < 1$, i.e

$$\begin{aligned} \lim{}_{n\to\infty} F^n(x) &= 0 \text{ if } x < x^F, \\ &= 1 \text{ otherwise.} \end{aligned}$$

The issue of degeneration's leads to the following theorem.

Theorem : Let $X_1, ..., X_n$ be a sequence of i.i.d random variables and let M_n be its maximum. If there exist two sequences of constants $a_n > 0$ and b_n such that

$$\lim P(\tfrac{M_n - b_n}{a_n} \leq x) = G(x) \text{ , as } n \to \infty,$$

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

-Gumbel :

$$\mathrm{G(x)} \;= e^{-e^{-x}}, \; -\infty \leq x \leq \infty$$

-Fréchet :

$$\begin{aligned} \mathrm{G(x)} &= e^{(-x)^{-\alpha}}, \text{ if } x > 0, \alpha > 0 \\ &= 1 \text{ otherwise;} \end{aligned}$$

-Reversed Weibull :

$$\begin{aligned} \mathrm{G(x)} &= e^{-(-x)^{-\alpha}}, \text{ if } x > 0, \alpha > 0 \\ &= 1 \text{ otherwise.} \end{aligned}$$

To those distributions can be added several parameters like the location parameter μ, the scale parameter σ by changing x to $\frac{x-\mu}{\sigma}$ and the shape parameter $\xi = \alpha^{-1}$. In the case of Fréchet or Reversed Weibull distributions, $\alpha \neq 0$, it is 0 for the Gumbel distribution. Together, they form the GEV distribution family where GEV stands for Generalized Extreme Value distributions. They share a common form, for which the parameters vary depending on the most suitable family :

$$F(x; \mu, \sigma, \xi) = e^{(-(1+\xi(\frac{x-\mu}{\sigma})^{-\frac{1}{\xi}})}, \text{ with } \mu \in \mathbb{R}, \xi \in \mathbb{R} \text{ and } \sigma > 0. \tag{1}$$

The type of distribution can be seen by the value of ξ, if $\xi = 0$ it is a Gumbel (type I) distribution, $\xi > 0$, a Fréchet (type II) distribution and if $\xi < 0$, a Reversed Weibull (type III) distribution.

On the other hand, considering only values above a specific threshold, this method constrasts with block-maxima which uses maxima in given time intervals. The main idea is to use groups of large values instead of a single one.
Let $X_1, ..., X_n$ be a sequence of i.i.d random variables having F as distribution function and let X be a random term from that sequence. Recalling Equation (1), if F meets the asymptotic requirements, then G as in (1) is GEV.

Now fixing a threshold u (suitable threshold being chosen with mean residuals life plots and POT plots), fitting the conditional distribution of the excedents $X - u$ given that $X > u$ can be calculated as :

$$P(X - u > x \mid X > u) = \frac{1 - F(u + x)}{1 - F(u)}.$$

As the distribution of F is unknown, the distribution of the threshold excedences is unknown too, see equation (1) and, using Taylor first order approximation of $log(F(z)) \simeq -(1 - F(z))$, if $F(x) \simeq 1$, then we have

$$\begin{aligned}\frac{1 - F(u + x)}{1 - F(u)} &\simeq (1 + \frac{\xi(\frac{x}{\sigma})}{1 + \xi\frac{(u-\mu)}{\sigma}})^{-\frac{1}{\xi}} \\ &= (1 + \xi\frac{x}{\hat{\sigma}})^{-\frac{1}{\xi}}, \text{ where } \hat{\sigma} \quad = \sigma + \xi(u - \mu).\end{aligned}$$

For the proof and details, see [**1**]. As said previously, equations (2.1) are only valid when $F(u) \simeq 1$ or if u is large enough with respect to the support of F. This is called the GPD, for Generalized Pareto Distribution.

2.2 Parameter estimation

The goal of the analysis is to find estimates for the parameters of those extreme value distributions GEV and GPD. Under the assumptions of independence and distribution, have

-GEV : First, also assuming $\xi \neq 0$, the log-likelihood function,

$$log(L(\mu,\sigma,\xi)) = -nlog(\sigma) - (1+\frac{1}{\xi})\sum_{i=1}^{n} log(1+\xi(\frac{x_i-\mu}{\sigma})) - \sum_{i=1}^{n}(1+\xi(\frac{x_i-\mu}{\sigma}))^{-\frac{1}{\xi}}$$

for $1+\xi(\frac{x_i-\mu}{\sigma}) > 0 \forall i \in [1,n]$

For $\xi = 0$, as in for Gumbel distributions, the log-likelihood function,

$$log(L(\mu,\sigma)) = -nlog(\sigma) - \sum_{i=1}^{n}(\frac{x_i-\mu}{\sigma}) - \sum_{i=1}^{n} exp(-(\frac{x_i-\mu}{\sigma}))$$

-GPD :

$$\begin{aligned} G_{\xi,\hat{\sigma}}(x) &= 1-(1+\xi(\frac{x}{\hat{\sigma}})^{-\frac{1}{\xi}}), \text{ when } \xi \neq 0; \\ &= 1-e^{-\frac{x}{\hat{\sigma}}}, \text{ when } \xi = 0. \end{aligned}$$

Recalling $\hat{\sigma}$ from (2.1), then let : $\hat{\sigma} > 0$ and $x \geq 0$ when $\xi \geq 0$ and $0 \leq x \leq -\frac{\hat{\sigma}}{\xi}$ when $\xi < 0$, giving the log-likelihood function for GPD,

$$\text{log(L}(\xi,\hat{\sigma})) = -nlog(\hat{\sigma}) - (1+\frac{1}{\xi}\sum_{i=1}^{n} log(1+(\frac{\xi x_i}{\hat{\sigma}})), \text{ for } \xi \neq 0.$$

When $\xi = 0$, the exponential case gives

$$\text{log(L))} \ = -nlog(\hat{\sigma}) - \sum_{i=1^n} 1 + \frac{x_i}{\hat{\sigma}}.$$

Solutions to these maximization problem are not analytical. Numerical solving is most of the time used.

2.3 Return periods estimates

Once the parameters have been estimated, one wants to calculate the return level estimates, i.e over a period $\frac{1}{p}$, such event has a probability of occurring of $1-p$. The longer the return period, the more likely to happen the event is.

-GEV : By inverting the extreme quantile of the GEV distribution 1 :

$$\begin{aligned} \text{x}_p &= \mu - \frac{\sigma}{\xi}(1-(-log(1-p))^{-\xi}), \text{ for } \xi \neq 0, \\ &= \mu - \sigma log(-log(1-p)), \text{ for } \xi = 0. \end{aligned}$$

-GPD : Similarly,

$$P(X > x \mid X > u) = (1 + \frac{\xi(\frac{x}{\sigma})}{1 + \xi\frac{(u-\mu)}{\sigma}})^{-\frac{1}{\xi}};$$

It follows that

$$P(X > x) = \zeta_u(1 + \xi(\frac{x-u}{\hat{\sigma}}))^{-\frac{1}{\xi}}, \text{ with } \zeta_u = P(X > u).$$

The return level x_p is the level that is exceeded on average once every p observations and is solution to

$$\zeta_u(1 + \xi(\frac{x_p-u}{\hat{\sigma}}))^{-\frac{1}{\xi}} = \frac{1}{p}.$$

2.4 Multivariate extreme value theory

Proceeding now with the multivariate extreme value theory.Let $X_n, n \geq 1$ be an i.i.d random vector in $\mathbb{R}$, and $X_k = (X_1^{(1)}, ..., X_k^{(d)}), k \in [1, n]$. The component-wise maxima is defined as :

$$\mathrm{M}_n = (M_n^{(1)}, ..., M_n^{(d)}) = (max_{k\in[1,n]}X_k^{(1)}, ..., max_{k\in[1,n]}X_k^{(d)}). \tag{2}$$

The interest here, resides in the asymptotic distribution of variable (2). Suppose then that $X_i = (X_1^{(1)}, ..., X_i^{(d)})$ have $F(X_1, ..., X_n)$ as distribution function and let

$$\begin{aligned} \mathrm{P}(\mathrm{M}_n \leq x) &= P(X_1 \leq x, ..., X_n \leq x) \\ &= F^n(x),\, x \in \mathbb{R}^d. \end{aligned}$$

The distribution of variable M_n,in equation (2), has a degenerate distribution as $M_n \to x^F$, where $x^F = sup\{x \mid F(x) < 1\}$. Under the assumption of existence of the normalizing sequences of constants $a_n^{(i)} > 0$ and $b_n^{(i)} > 0$ for every $i \in [1, d]$ and $n \geq 1$ such that

$$P(\frac{M_n^{(i)} - b_n^{(i)}}{a_n^{(i)}}) \leq x^{(i)}, i \in [1, d]) = F^n(a_n^{(1)}x^{(1)} + b_n^{(1)}, ..., a_n^{(d)}x^{(d)} + b_n^{(d)} \to G(X^{(1)}, ..., X^{(d)}).$$

The limiting distribution G has each marginal distribution G_i for $i \in [1, d]$ being non-degenerate. The i^{th} marginal distribution is

$$\mathrm{F}_i^n((a_n^{(i)}x^{(i)} + b_n^{(i)}) \to G_i(X^{(i)}).$$

From univariate results equation (1), each G_i is a member of the GEV family.

2.5 Bivariate extreme value distributions

The most frequently used multivariate extreme value distributions are the bivariate ones. Starting with a definition :

Definition : $G(x)$ is max-stable if for every $i \in [1,d]$ and every $t > 0$, there exist functions $\alpha^{(i)}(t)$ and $\beta^{(i)}(t)$ strictly positives such that

$$\mathrm{G}^t(x) = G(\alpha^{(1)}(t)x^{(1)} + \beta^{(1)}(t), ..., \alpha^{(d)}(t)x^{(d)} + \beta^{(d)}(t)).$$

It can be shown that $G(x)$ is max-stable if and only if it is a multivariate extreme value distribution. Thus one needs to find all possible multivariate max-sable distribution, assuming one of the three possible univariate marginal extreme valued distribution. Also can be shown that any bivariate extreme value distribution with unit Fréchet margins can be written as

$$\mathrm{G}_*(x,y) = e^{(-(\frac{1}{x}+\frac{1}{y})A(\frac{x}{x+y}))}, \tag{3}$$

where $A(\omega)$ is called the dependence function. Since (3) has unit Fréchet margins, we have

$$\lim_{x\to\infty} G_*(x,y) = e^{-\frac{1}{y}} \text{ and}$$
$$\lim_{y\to\infty} G_*(x,y) = e^{-\frac{1}{x}} \,;$$

with

$$\mathrm{G}_*^n(x,y) = G_*(\frac{x}{n}, \frac{y}{n}).$$

Implying that $G_*(x,y)$ is max-stable. It can also be shown that $G_*^t(xt, yt) = G_*(x,y)$, for $t > 0$.

It can be shown that the dependence function $A(\omega)$ has the following properties :

1. $A(0) = A(1) = 1$;
2. $max(\omega, 1-\omega) \leq A(\omega) \leq 1$, if $0 \leq \omega \leq 1$;
3. $A(\omega)$ is convex for $\omega \in [0,1]$.

A has lower bound

$$\mathrm{A}(\omega) = 1 - \omega, \text{ for } \omega < \tfrac{1}{2},$$
$$= \omega, \text{ otherwise,}$$

and upper bound $A(\omega) = 1$.
There is no parametric family which gives all possible bivariate extreme value distribution.

Starting there with the R-package *evd*, there are 9 different parametric bivariate extreme value models. Listed in the following subsection. One first needs a finite form for the marginal distributions :

$$\mathrm{y}_i = y_o(x_i) \quad = \{1 + \xi_i(\frac{x_i - y_i}{\sigma_i})\}^{-\frac{1}{\xi_i}}, \text{ for } i = 1, 2$$

where the marginal parameters are (μ_i, σ_i, ξ_i), with $\xi_i > 0$.
If $\xi_i = 0$, y_i is defined by continuity.

In each of the 9 parametric bivariate distribution functions G given below, the univariate margins belong to the GEV family as explained earlier.
Choosing an appropriate block-size for the block-maxima as it is the kind of model we want to fit to our data.

2.6 Bivariate block-maxima

We will proceed the analysis using a bivariate model for erosion and oriented wind speed. Due to the format of the data available, we will use component-wise block maxima models, with as before, blocks of one year (starting January, the 1^{st} until December, the 31^{st}).

Suppose there is $(X_1, Y_1), (X_2, Y_2), ...$ a sequence of vectors that are independent versions of a random vector (X, Y) having distribution $F(x, y)$, then define as previously the component-wise maxima :

$$\begin{aligned}\mathrm{M}_n &= (max_{i=1:n}(X_i), max_{i=1:n}(Y_i)) \\ &= (M_{x,n}, M_{y,n})\end{aligned}$$

Note that it is a component-wise maxima vector, i.e, this vector does not have to be an observed vector from the original series of data. The two maxima may have been observed at different moments in the block.

— log : The bivariate logistic distribution function is defined as :

$$\text{G(x,y)} \quad = exp\{-\{x^{\frac{1}{r}} + y^{\frac{1}{r}}\}^r\},$$

where $r \in [0,1]$. Independence when $r = 1$, dependence when r tends to 0. This is a special case of the alog model (following model).

— alog : The bivariate asymetric logistic distribution function is defined as :

$$\text{G(x,y)} = exp\{-(1-t_1)x - (1-t_2)y - \{(t_1x)^{\frac{1}{r}} + (t_2y)^{\frac{1}{r}}\}^r\},$$

where $r \in [0,1]$, $t_1 \geq 0$ and $1 \geq t_2$. Independence is either when $r = 1$ and $t_1 = 0$ or $t_2 = 0$. Dependence when r tends to 0 and $t_1 = t_2 = 1$.
This is the origin of the special case of the log model (previous model), as in where $t_1 = t_2 = 1$. Different limits occur when t_1 and t_2 are fixed and r tends to 0.

— hr : The Hustler-Reiss distribution function is defined as :

$$\text{G(x,y)} = exp\{-x\phi(\frac{1}{r} + r(log\frac{x}{y})) - y\phi(\frac{1}{r} + r(log\frac{y}{x}))\}$$

where ϕ is the standard normal distribution function and $r > 0$. Independence is obtained when r tends to 0. Dependence when r tends to ∞.

— neglog : The bivariate negative logistic distribution function is defined as :

$$\text{G(x,y)} = exp\{-x - y + (x^{-r} + y^{-r})^{-\frac{1}{r}}\}$$

where $r > 0$. Independence is when r tends to 0 and dependence when r tends to ∞. This is a special case of the aneglog model (following model).

— aneglog : The bivariate asymmetric negative logistic distribution function is defined as :

$$\text{G(x,y)} = exp\{-x - y + ((t_1x)^{-r} + (t_2y)^{-r})^{-\frac{1}{r}}\}$$

where $r > 0$, $t_1 \geq 0$ and $1 \geq t_2$. Independence is when either r,t_1 or t_2 tends to 0 and dependence when $t_1 = t_2 = 1$ and r tends to ∞.
This is the origin of the special case of the neglog model (previous model), as in where $t_1 = t_2 = 1$. Different limits occur when t_1 and t_2 are fixed and r tends to 0.

— bilog : The bivariate bilogistic distribution function is defined as :

$$\text{G(x,y)} = exp\{-\{xq^{1-\alpha} + y(1-q)^{1-\beta}\}^r\},$$

where q is the root of the following equation :

$$\text{(1-}\alpha\text{)}x(1-q)^\beta - (1-\beta)yq^\alpha = 0$$

where $0 < \alpha$ and $\beta < 1$. Independence when $\alpha = \beta$ approaches 1. Dependence when $\alpha = \beta$ tends to 0. Different limits occur when α or β is fixed and the other tends to 0. This is a special case of the log model $(g.1)$ when $\alpha = \beta$ and then to the alog model $(g.2)$ with also $t_1 = t_2 = 1$.

— negbilog : The bivariate negative bilogistic distribution function is defined as :

$$\text{G(x,y)} \ = exp\{-x - y + xq^{1+\alpha} + y(1-q)^{1+\beta}\}^r\},$$

where q is the root of the following equation :

$$(1{+}\alpha)xq^{\alpha} - (1+\beta)y(1-q)^{\beta} = 0$$

where α,$\beta > 0$. Independence when $\alpha = \beta$ approaches ∞. Dependence when $\alpha = \beta$ tends to 0. Different limits occur when α or β is fixed and the other tends to 0. This is a special case of the neglog model $(g.1)$ when $\alpha = \beta$ and with reformulation $\frac{1}{\alpha}$ and $\frac{1}{\beta}$.

— ct : The Coles-Tawn distribution function is defined as :

$$\text{G(x,y)} \ = exp\{-x(1 - Be(q;\alpha+1,\beta)) - yBe(q;\alpha,\beta+1)\}, \text{ where}$$
$$\alpha,\beta < 0 \text{ and } q = \frac{y\alpha}{y\alpha + x\beta}.$$

$Be(q;\alpha,\beta)$ is the Beta distribution function evaluated at q. Independence is when $\alpha = \beta$ approaches 0 or when one of α, β is fixed and the other tends to 0. Dependence is when $\alpha = \beta$ tends to ∞. Different limits occur when one of α, β is fixed and the other tend to ∞.

— amix : The asymmetric mixed distribution function is defined as :

$$\text{A(t)} \ = 1 - (\alpha+\beta)t + \alpha t^2 + \beta t^3$$

where $\alpha \geq 0$ and $+3\beta \geq 0$, or where $\alpha + \beta \leq 1$ and $\alpha + 2\beta \leq 1$. Then,$\alpha \in [0, 1.5]$ and $\beta \in [-0.5, 0.5]$.
Although $\alpha \to 1$ implies $\beta < 0$.
Independence is when $\alpha = \beta = 0$. Dependence is when α increases and β is fixed, although complete dependence can't be reached.

2.7 Copula Theory

A bivariate model is needed for the wind speed and the erosion, and for such, the copula theory provides an alternative way of creating models, i.e. : Extreme value copula models. Below, a summary of the copula theory is provided but refer to [**2**] for more information on this topic.

The main idea of copulas can be explained as the following : When the joint distribution function is hard to find or to use, copulas are a way of having the joint distribution as a function of the marginal distributions. Later on in this report using a copula and the two marginal distributions a model for oriented windspeed and erosion will be created.
Let (X, Y) coming from a $F(x, y)$ distribution, with $F : \mathbb{R}^2 \to [0, 1]$. And define $C : [0, 1]^2 \to [0, 1]$.

$$F(x, y) = C(F_1(x), F_2(y)),$$

with F_1 and F_2 the respective marginal distributions of the random variables$X$ and Y.

Then, a d-dimensional copula is defined for $d \geq 2$ such that it is a function $C : [0, 1]^2 \to [0, 1]$ for which

$$\exists (U_1, ..., U_d) \text{ such that, } U_i \sim U(0, 1) \forall i \in [1, d],$$

and

$$C(u_1, ..., u_d) = P(U_1 \leq u_1, ..., U_d \leq u_d).$$

Also, the copula associated with F is defined for $F : \mathbb{R}^d \to [0, 1]$,

$$C(u_1, ..., u_d) = F(F_1(u_1)^{-1}, ..., F_d(u_d)^{-1}),$$

where $F_1, ..., F_d$ are the marginal distributions of F.

The next step is the most important theorem about copulas, Sklar's theorem

Theorem : Let F be a joint distribution function with marginals $F_1, ..., F_d$. Then, there exists a copula C such that

$$F(X_1, ..., X_d) = C(F_1(X_1), ..., F_d(X_d))$$

And conversely, if C is a copula and $F_1, ..., F_d$ are distribution functions, then it is a joint distribution function with marginals $F_1, ..., F_d$.

There are three other important consequences concerning the last theorem :

- C explains the dependence between the margins ;
- a joint distribution functions can be split in two parts, its margins and its copulas. Hence, they can be modelled separately ;

— for a given copula C, the margins can be freely changed, for example, $C(H_1(X_1), ..., H_d(X_d))$ is a proper distribution function.

Aside from Sklar's theorem, it can be shown that any measure of dependence which depends only on a copula, does not change under strictly increasing transformations. Which gives us Kendall's τ, and Spearmann's ρ dependence measures.

An extreme-value copulas is then defined as :

Definition : Any copula for which $C^t(u,v) = C(u^t, v^t), \forall t > 0$, is called an extreme-value copula. This also applies to any dimension $d \geq 2$, where $C^t(u_1, ..., u_d) = C(u_1^t, ..., u_d^t)$.

It can be shown using a bivariate GEV with unit Fréchet margins G_* that

$$C(u,v) = exp(log(uv)A(\frac{log(u)}{log(uv)}),$$

with A a convex function called dependence function and u, v in $[0,1]$.

From that definition, can be found the different extreme value copulas like the Gumbel, Galambos, Tawn, t and Husler-Reiss.

2.8 Model validation

As one creates models to fit the data, one also needs to check if the fit is good enough. In this analysis, the "fit diagnostic plot" function from the *in2extRemes* R package will be mostly used for the univariate part. But one can also use various criteria such as $p - value, AIC$ and BIC, or dependence functions for bivariate models.

Definition : The Akaike Information Criteria (AIC) is commonly use for model order selection. The value of AIC is based on information theory, rewarding a high likelihood of the model but penalizing an high model order. The lower the AIC is, the better the model is (but it is NOT a sign that the model is good, just that it is better than other models with a higher AIC).

$$AIC\ = -2log(L) + 2k,$$

with L the likelihood function and k the number of parameters in the model.

Definition : The Bayesian Information Criteria (BIC), works like the AIC above. A lower

BIC indicates a better model (same note as for AIC about how good the model is, BIC is not a quality indicator but an efficiency one)

$$BIC \ = -2log(L) + log(n)k$$

with L the likelihood function, n the number of observations and k the number of parameters in the model.

Those information criterion, AIC and BIC are not a measure of goodness of fit and one doesn't penalize over-fitting(AIC) where the other doesn't penalize under-fitting(BIC). They are indicators of which models is the best but not if a model is good.

Fit Diagnostic : Let $x_1, ..., x_n$ denote a sample of i.i.d observations with distribution F. Have $\hat{F}$, an estimate of F. The empirical function is defined by

$$\tilde{F}\textit{(x)} \ = \frac{i}{n+1}, \textit{ for } x \in [x_{(i)}, x_{(i+1)}].$$

As $\tilde{F}$ is an estimate of the true probability distribution function F, it should be similar to the estimated model $\hat{F}$. Comparing $\hat{F}$ with $\tilde{F}$ shows various goodness of fit procedures, where probability plot and quantile plot are the two most commonly used graphical techniques.

- *A probability plot is the set of points $((\hat{F}(x_{(i)}), \frac{i}{n+1}), i \in [1, n])$;*
- *A quantile plot is the set of points $((\hat{F}^{-1}(\frac{i}{n+1}), x_{(i)}), i \in [1, n])$;*

If $\hat{F}$ is a good enough estimate, the probability plot and the quantile plot will be close enough from the diagonal unit line. Opposed to the AIC and BIC, those graphical methods for goodness of fit show whether a model is good enough, but doesn't say anything on it's efficiency about the number of parameters.

2.9 Dependence function for bivariate models :

Now going to go quickly through several methods of estimating the dependence function, which is a good way of assessing the goodness of fit of a model onto the data. Starting with Non-Parametric estimation of dependence function.

Suppose $(X, Y) \sim G_*(x, y)$, a couple of random variables and $(x_1, y_1), ..., (x_n, y_n)$ a sample ;

I).Non-parametric methods :

i). Pickand's estimator(1981) :
It can be shown that $min(\frac{1}{(1-\omega)X}, \frac{1}{Y\omega}) \sim exp(A(\omega))$, where $A(\omega)$ is the dependence function. Defining $Z_i(\omega) = min(\frac{1}{(1-\omega)X_i}, \frac{1}{Y_i\omega})$, one can get :

$$\hat{A}_p(\omega) = \frac{n}{\sum_{i=1}^{n} z_i(\omega)}$$

Although, even if this estimator is good asymptotically, the reality of the data makes this estimator quite inaccurate. There is a modified version of this estimator defined later on.

ii). Capéraà, Fougères and Genest's (CFG) estimate (1997) :
Take $(U_i, V_i) \sim (F_1(X_i), F_2(Y_i)), \forall i \in [1, n]$.
The CFG estimator of the dependence function is :

$$\begin{aligned} A_n(\omega) &= (1-\omega)Q_n^{1-p(\omega)}, && \text{if } \omega \in [0, Z_{(i)}] \\ &= \omega^{i/n}(1-\omega)^{1-i/n}Q_n^{1-p(\omega)}Q_i^{-1}, && \text{if } \omega \in [Z_{(i)}, Z_{(n)}] \\ &= \omega Q_n^{-p(\omega)}, && \text{if } \omega \in [Z_{(n)}, 1]. \\ \text{Where } Q_i &= \{\Pi_{k=1}^{i} \frac{Z_{(k)}}{(1-Z_{(k)})}\}^{\frac{1}{n}}, && \forall i \in [1, n]. \end{aligned}$$

II). Maximum likelihood based on parametric models :

iii). Hall-Tajvidi modification of the Pickand's estimator :
By changing $\hat{X}_i = \frac{1}{X_i}$ and $\hat{Y}_i = \frac{1}{Y_i}$, have

$$\hat{B}_{p-HT}(\omega) = \frac{\sum_{i=1}^{n} min(\frac{\hat{x}_i}{(1-\omega)E[\hat{X}]}, \frac{\hat{y}_i}{(1-\omega)E[\hat{Y}]})}{n}$$

iv). Constrained smoothing splines :

$\hat{A}$ can be approximated by a spline that is constrained to satisfy all the necessary conditions of the dependence function, bu choosing regularly spaced points from $t_0 = 0$, to $t_m = 1$, all points being strictly smaller than the next one, spanning the interval $[0, 1]$.

Then, with $s > 0$, one has to find $\tilde{A}_s$, the polynomial of degree 3 or more that minimizes :

$$\sum_{j=1}^{m}(\hat{A}(t_j) - \tilde{A}_s(t_j))^2 + \int_0^1 \tilde{A}''{}_s(t)^2 dt \tag{4}$$

As mentioned previously, this function satisfies the dependence function requirements.

For more details about the theory behind stationary stochastic processes, see [**8**] ; for extreme value theory, see [**1**] ; finally, and for copula theory, see [**2**].

3 Data analysis

Data sets consist of yearly maxima for both wave runups and dune erosion, then hourly measurements for sea levels at Viken.
Around the coast of Ängelholm, there are three stations collecting wind speeds and wind direction for the SMHI (Swedish Meteorology and Hydrology Institute), Kullen, Hallands Väderö A and Barkakra. They are located as shown in the Figure 3.

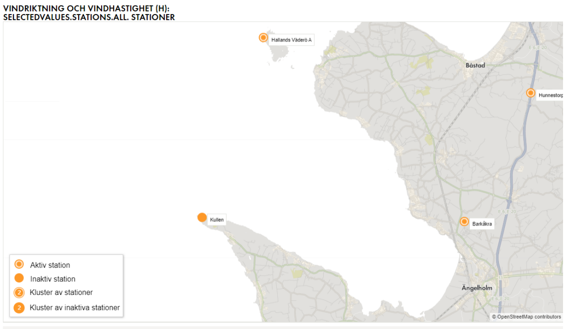
VINDRIKTNING OCH VINDHASTIGHET (H): SELECTEDVALUES.STATIONS.ALL. STATIONER
Hallands Väderö A
Båstad
Hunnestorp
Kullen
Barkåkra
Ängelholm
Aktiv station
Inaktiv station
Kluster av stationer
Kluster av inaktiva stationer
© OpenStreetMap contributors

Figure 3 – Coast of Ängelholm with the three stations of interest.

The data coming from the SMHI were given with an accuracy mark depending if it was measured (verified) or computed by one of their models (unverified). Later on, there will be a discussion about using verified and/or unverified data for the analysis.

The wave runups and the dune erosion data were given at 40 different spots of the shore, numbered from 1 to 40. When discussing a specific spot on the shore and the data from there, it will be called "Lats", short for "Location At The Shore". For example, point 2 Lats is the second part of the shore from the North.
As the period of interest belongs in between the years 1976 – 2015 (data from dune erosion and maximum run-ups where collected along that period), a few problems appeared with the available data. They are discussed below.

3.1 Incomplete data

The Kullen station, located on the south-west of the coast, at its very end, only had data collected until 1996 or so, which represent half of the period. Then, the Hallands Väderö A station only had data available from 1996 to 2016, the other half of the period. Finally, Barkakra had data available from 1976 to 2002 and from 2008 to 2016, so about 35 out of the total 40 years wanted.

Also, the yearly maximum erosion has many of 0-values which makes the analysis more difficult.

3.2 Data preparation

As parts (often very consequent ones) of the data are missing, it is required to find a way to get around. The following list consists of the various ideas of data preparation that was, or could have been, used in this analysis ; they come with different advantages as well as disadvantages. Some of them use, for example (b.1) and (b.2) use the data as available, i.e reduced data sets, where the other alternatives are ways to get a more complete set but that may require strong assumptions that can be hard to fill in correctly.
They are as follow :

— b.1 : Only using 20 years of data instead of 40, then, one can either use the data from Kullen or Barkakra for the first 20 years, or from Hallands Väderö A for the last 20 years, depending on which period is the most interesting.

Problem (b.1) :
The period is somewhat short and if a block maxima model appears to be the best, the accuracy will decline due to low amounts of data. Also, as the previous study was made on the whole 40 years, it would be better to aim for consistency ;

— b.2 : Barkakra's set has the most years available, but as they are not consecutive, they may not be usable as they are. Although, checking whether the set is stationary enough would make it usable.

Problem (b.2) :
It would still be 35 years instead of 40 years, lowering the accuracy of the results ;

— b.3 : Have a mix of them, i.e use 20 years of one and 20 years of the other, which, grouped, would make a whole 40 years data set as they conveniently cover each other's gap. There is even have a choice between Barkakra and Kullen for the first part, choosing the one that shows the most correlation with Hallands Väderö A to get a significantly close set. As the stations are part of the same coast, their properties could be similar.

Problem (b.3) :
They are not. Sets that are significantly close are needed. Especially that, looking at the data, one can see that the wind speeds are very different (much higher) at Hallands Väderö A than at the others. Mixing the data sets as proposed might not

be relevant ;

A mix of the 3 sets seems rather inefficient, as it would be : 20 years from Kullen (1976 – 1996), 6 years from Barkakra (1997 – 2002), 5 years from Hallands (2003 – 2008) and the final years from Barkakra again (2009 – 2015). This case is positively useless as it takes the flaws of pretty much all other combinatory alternatives and doesn't gain any of its advantages other than having a full 40-years set.

— $b.4$: A different mixing could be done by using the set from Barkakra and using 5 years from another station (Hallands Väderö A for example) to cover its gap. As Barkakra is the set with the most usable years by far, it would make more sense than the previous two alternatives ($b.2$) and ($b.3$). It also is the closest station from the actively eroded zones Lats.

  Problem ($b.4$) :
  Same as ($b.3$).

— $b.5$: Then, rather than directly using data from Hallands Väderö A to covers Barkakra's gap, one could simply reconstruct Barkakra's missing 6 years by creating a model that fits Barkakra's data. Either by using Hallands Väderö A station to help (like in a Box-Jenkins model, see [**7**]) or without, by using a Kalman filter to make the reconstruction, knowing that wind speeds can be modelled via a Weibull distribution.

  Problem ($b.5$) :
  Such a reconstruction might be too much of a hassle and give questionable results, where alternative ($b.4$) is simpler but also seems efficient enough. Using such methods to re-create a maximum of 5 points (approximately from the years 2002 – 2008) appears tedious, at best.

Other than the faulty wind data, the yearly maxima for erosion show values and behaviours that are not easy to use. Most of the data points are 0-valued, making the analysis intricate. By grouping them or using blocks of several years (no more than two preferably), the erosion data becomes more usable even though the number of data depletes fast.

— $b.6$: Instead of using 40 years of data, i.e 40 yearly maxima, one could use half (20) and take a block size of two-years instead of one.

  Problem ($b.6$) :

Having 20 points instead of 40 may be too small an amount for the analysis to work properly.

- $b.7$: As a way of minimizing the faulty erosion data, instead of choosing one spot Lats, one could pick three neighbouring spots Lats and take the yearly maximum over those 3 ones. As they are close to each other, the values should be concordant and somewhat take care of most of the 0s. The ad hoc spots would then be spots 9 to 11 Lats, but still represent 10 different 0-valued points.

  Problem ($b.7$) :
  Are they close enough ? If they are, the erosion being 0 somewhere might imply the erosion would be 0 on the other points too. It also represent a quarter of the whole dataset being 0, which is too much.

- $b.8$: Taking the maximum erosion over the whole bay (all spots Lats) so the chances of having 0s should be significantly reduced.

  Problem ($b.8$) :
  It actually is not. The 0-values on yearly erosion seem to be unrelated to the spot Lats but related to specific years. Using a maximum along every spot Lats doesn't solve that problem.

  Following from last approach, ($b.8$), a block size of two years will be taken for erosion over the whole bay . Calculations in the next part show that the maxima are actually located around spots Lats 9 to 11, hence that a maxima over the whole bay is the same as a maxima over those three spots Lats.

As a final note on those alternatives, trying to get data on a shifted period, as in $1979-2019$ for all the data sets (wind speeds, wind directions, erosion, wave run-ups and sea levels) would not be of any help at solving the above problems.

This paper will only contain the most successful attempt and a comparison between the models created with this approach. That means the data processing that will take into account the smallest confidence intervals, the significance of the parameters and the fit to the data for example.
For this analysis, the alternatives ($b.4$), ($b.6$) and ($b.8$) will be used, i.e, filling Barkakra's data set gap with Hallands Vder data and using the maximum erosion over the whole bay. The amount of manipulations done on the data was deemed necessary to get functioning models, the other attempts were only ending in insignificant models.

3.3 Verified and Unverified data

To add to the previous section on missing data, there is a certainty, or accuracy problem.

The wind data as they were collected and stored were given the status "Verified (G)" or "Unverified(Y)", as "G" means "kontrollerade och godknda vrden" (or "checked and approved values") where "Y" means "misstnkta eller aggregerade vrden. Grovt kontrollerade arkivdata och okontrollerade realtidsdata (senaste 2 tim)" (or "suspicious or aggregated values. Roughly controlled archive data and uncontrolled real-time data (in the last 2 hours)").

Also with it came some data that had G-valued wind speeds but Y-valued wind directions (or vice-versa). One can, either consider such an event as a Y-valued data vector if one uses both wind speed and wind direction, or as G-valued if one only uses the G-valued component data in the vector. Although, both components will be used during the analysis, and therefore, the whole vector is to be referred as Y-valued.

By checking the yearly maxima of those data sets, it revealed that a significant amount of them were stated in the Y category, hence were unverified. 50% of Hallands Väderö's set for yearly maxima contains unverified values. The same ratio is true for the other sets. As it is half of them being unverified, one can wonder if running an analysis with that many uncontrolled values is worth anything.

It is now a question of whether to use all the data or only the verified ones. Although, for the latter, another problem would rise as the frequency of apparition of those unverified data may not be from a special or predictable pattern, hence there would be years with less values than others. Depending on the analysis, that can be problematic.

As a consequence of not wanting to have different amounts of data in each year, and, more importantly, even if the data is unverified, it can be the true value or significantly close to the true value. As the quality of the models made by SMHI cant be judged, the data was kept, regardless of verified status.

4 Extreme value modelling

As two of our data sets (dune erosion and wave run-up) are yearly maxima, they are best fitted for the block maxima models. Trends and seasonal cycles can't be seen from any plot in the data because of its nature. Both data sets consists of 59 points spread homogeneously along the shore with 40 yearly maxima (years 1976 – 2015).

The block size is chosen from the available data, that is that the blocks should be cut in years, from the 1^{st} of January to the 31^{st} of December of that year. Although, for the bivariate model between dune erosion and oriented wind speed, a lack of non-0 values will show up and create problems with the model; hence a change needs to be made and a maximum over two-years sized blocks instead of one-year sized blocks is used in order to minimize their impact on the model.

A significance level of 5% was chosen to concur with results from other studies and make it easier to compare as it is the most commonly used. All univariate block-maxima models have been tested for GEV distribution against Gumbel distribution. For peaks over threshold, models with shape = 0 are tested against exponential distributions. Whatever distribution fitting the most the data will be kept according to likelihood-ratio-tests, and the other rejected. Unless specified otherwise, all confidence intervals will be calculated using normal approximation from the R-package *in2extremes*.

4.1 One-dimensional analysis of the data

The first two parts of this section will be dealing with sea-level rise and wave runups. The two of them together, by simple addition of the return levels will give the potential risk of flooding, i.e, if the wave height leveled up by the sea level, bypass the sand dunes height, there is a risk of flood.

The last part of this section will be dealing with the erosion, and wind components, speed and direction. A univariate analysis will be performed to first understand how wind and erosion behave on their own and then in the later multivariate analysis section to confirm the effect of wind on dune erosion.

On to sea level

The sand dune is the main protection against flooding. As long as the sea-level or the wave height is lower than the dune height, coastal cities and lands will not be damaged. For an in depth analysis of the dune themselves, see [**3**], Long-term beach and dune evolution, 2019.

Here, only use extreme value analysis will be used in regards to the sea-level and then on to wave runups, rather than an analysis on the global shape and parts of the dune and how both sea level and wave height impact them.

Figure 4 is a scatter plot of the sea level above or under 0 collected at the station Viken, in the bay just south to Ängelholm. Data in centimeters.

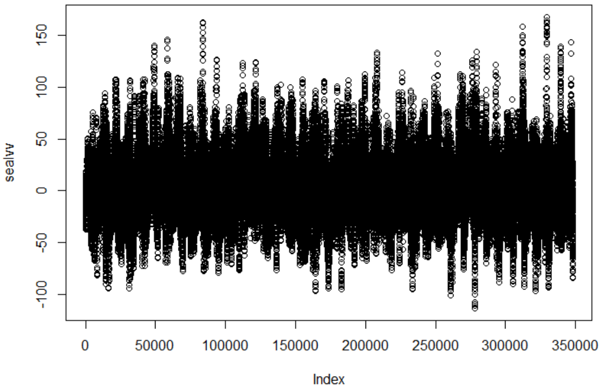

Figure 4 – Sea level scatter plot against index (1976-04-22 to 2015-12-31) in Viken.

The plot shows that there is some kind of cyclic behaviour in the data, upholding very high levels and very low alike. Most of the highest sea levels recorded are from the winter period, i.e. November to February. Therefore, only these winter periods were chosen for the study.

As data is an hourly measurement, the set still contained too many values which were reduced by taking the daily maxima, see Figure 5. The interest here, is to perform a peak over threshold analysis, using GPD models. In this way, part of the data is already declustered, taking care of the hourly, if not daily, dependence in the values.

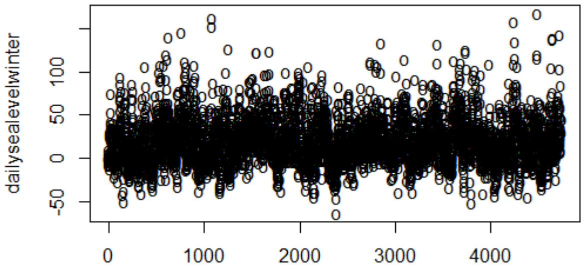

Figure 5 – Scatter plot for daily maxima winter sea level in Viken ; water level is given above or under 0, ground level.

The threshold, chosen, using mean-life residual plot and POT plot functions could be located between 70 and 90 centimeters above ground 0. Fitting several models with thresholds from this interval, two conclusions were drawn :

— the shape parameter was never significant ;
— a threshold of 70 centimeters seems to fit the data best.

Among three different thresholds, respectively 70, 80 and 90 centimeters above ground zero, the first one was deemed more satisfactory. This model had the best fit on diagnostic plots, and lowest confidence interval spread for the parameters.
Even though the other two models had their strengths, such as a lower AIC and BIC, a narrower confidence interval's spread on the return levels or higher MLE, they were rejected, mainly based on goodness of fit.
Table 1 contains parameter estimates for the sea level.

Table 1 – Sea level parameter estimates with GPD distribution ; number in brackets are 95% confidence intervals for the parameters.

scale σ	shape ξ
21.20 (16.83, 25.57)	-0.04 (-0.19, 0.11)

Shape was deemed non-significant and thus set to 0 after using profile likelihood on the estimates.

Table 2 contains 10−,100− and 200-year return level with 95% confidence intervals.

Table 2 – Sea-level return levels for 10-, 100- and 200-year return periods with GPD ; number in brackets are 95% confidence intervals for the return levels.

10-year return level	100-year return level	200-year return level
106.80 (142.02, 191.59)	204.73 (150.58, 258.89)	215.47 (150.44, 280.51)

Return-levels have relatively quickly growing confidence interval's spreads but nothing alarming. Figure 6 is a diagnostic plot for the sea level at Viken fitted with GPD using a threshold of 70 centimeters above 0. Also, just by looking at the numbers, one can be alarmed by the possibility of rising that the sea level can be seen as it seems to be possibly increasing more than 2 meters.

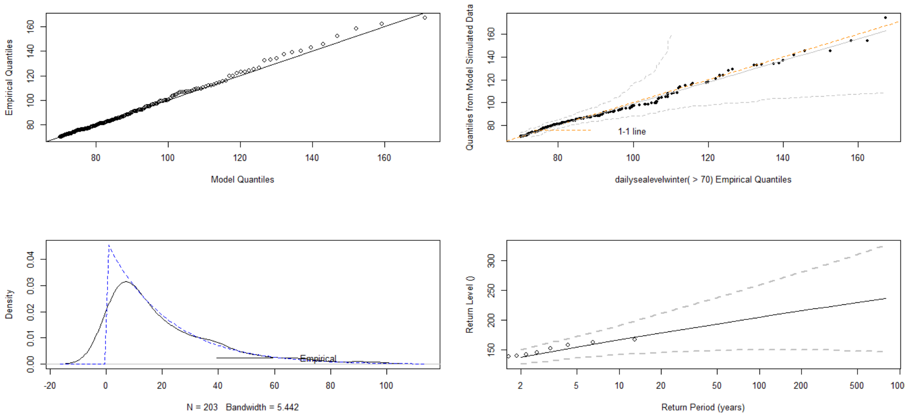

Figure 6 – Diagnostic plot for sea level at Viken with Generalized Pareto Distribution (GPD)

The diagnostic plot in Figure 6 is satisfactory. The conclusion is the same for empirical quantiles against modeled ones. The density function and return level plots are good enough.
This GPD model with threshold $u = 70$ cm and shape parameter at 0 seems to fit the data well enough.

Although, as said previously, the shape parameter ξ is not significantly different from 0 (result checked by normal approximation and profile likelihood methods), it can be set to 0, the GPD then becomes a standard exponential distribution.

The data then would have an exponential distribution with parameter :$\sigma = 20.37$ and a standard deviation of 1.43. Note that this interval is contained in the previous confidence interval proposed by the GPD model which with a likelihood ratio test leads us to rejecting the exponential model with a probability of 95%.

Later on, in the conclusion part, some interpretation of the results will be done considering flood risks in the bay of Ängelholm.

On to maximum run-up

Figure 7 is a scatter plot of the yearly maxima for the run-up over the period (1976 – 2015) on Ängelholm's shore at point Lats (Location At The Shore) 1. Data in meters.

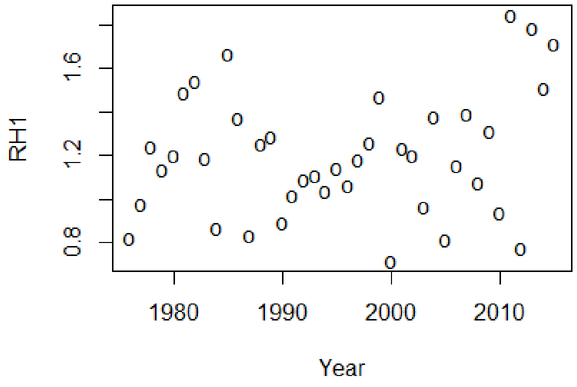

Figure 7 – Yearly maxima for run-up scatter plot from point 1 Lats

Three points Lats were chosen to be looked at as it would be tedious and worthless to analyse all 59.

One can see, looking at some scatter plots between two points on the shore that there is a very strong (and positive) correlation between the points. Especially from one point to the following one, where the correlation neighbours 1 almost all the time. Their dependence then seems quite obvious, hence, taking points as far away from each other as the points 1, 25 and 59 should mitigate the problem to the highest degree as possible. Although, this is not true between points 1 and 2. As much as there is a positive correlation there, it is significantly less strong as for the others which can be due to its position Lats. Then, point 2 will be looked at during this block maxima analysis which seems to be more significant regarding its values.

Using likelihood-ratio test, the Gumbel model couldn't be rejected in the four spots of the coast (1, 2, 25 and 59). The following tables contain parameter estimates for locations and scales as well as various return levels.

Table 3 contains parameter estimates at the indicated locations along the shore.

Table 3 – Maximum run-up with Gumbel distribution ; Lats stands for Location along the shore ; number in brackets are 95% confidence intervals for the parameters.

Lats	location μ	scale σ
1	1.07 (0.99, 1.14)	0.24 (0.18, 0.30)
2	1.60 (1.50,1.71)	0.32 (0.25, 0.40)
25	1.58 (1.48, 1.69)	0.31 (0.24, 0.39)
59	1.35 (1.26, 1.44)	0.28 (0.21, 0.35)

Table 4 contains 10−,100− and 200-year return level with 95% confidence interval for those points along the shore.

Table 4 – Maximum run-up with Gumbel distribution ; Lats stands for Location along the shore ; number in brackets are 95% confidence intervals for the return levels.

Lats	10-year return level	100-year return level	200-year return level
1	1.60 (1.43, 1.77)	2.16 (1.87, 2.46)	2.33 (1.99,2.67)
2	2.33 (2.10, 2.56)	3.09 (2.69, 3.49)	3.31 (2.86, 3.77)
25	2.29 (2.06, 2.52)	3.03 (2.63, 3.43)	3.25 (2.80, 3.70)
59	1.98 (1.78, 2.18)	2.64 (2.29, 2.99)	2.834 (2.43, 3.23)

Notwithstanding the fact that the four points belong to the same shore and therefore are not far away from each other, there still are differences in their estimates.

The parameter's estimates for the points 2, 25 and 59 fall into each other's confidence intervals where 1's barely does the same. Studying the spread of the confidence intervals could indicate the most accurate point to study wave run-ups. In that case, it would be spot number 1. Although, from the data and the return levels in Table 4, one can see that they are way lower than the others.

Just as in the previous part, exception made for spot 1 along the shore which estimates and confidence intervals still don't cross the other's. This could come from the very position of that spot as it is the first one on the shore and may be protected by the shape of the coast itself. Run-ups there seems to be smaller. As the intention is to model the maximum risks of flood on the coast of Ängelholm, point 1 should not be taken as the main example there. Dealing with Gumbel distribution (shape $\xi = 0$), the MLE of the upper end-point is infinity.

Figure 8 shows a diagnostic plot for the data from point 2 fitted with a Gumbel distribution from which the previous estimates and confidence intervals where taken.

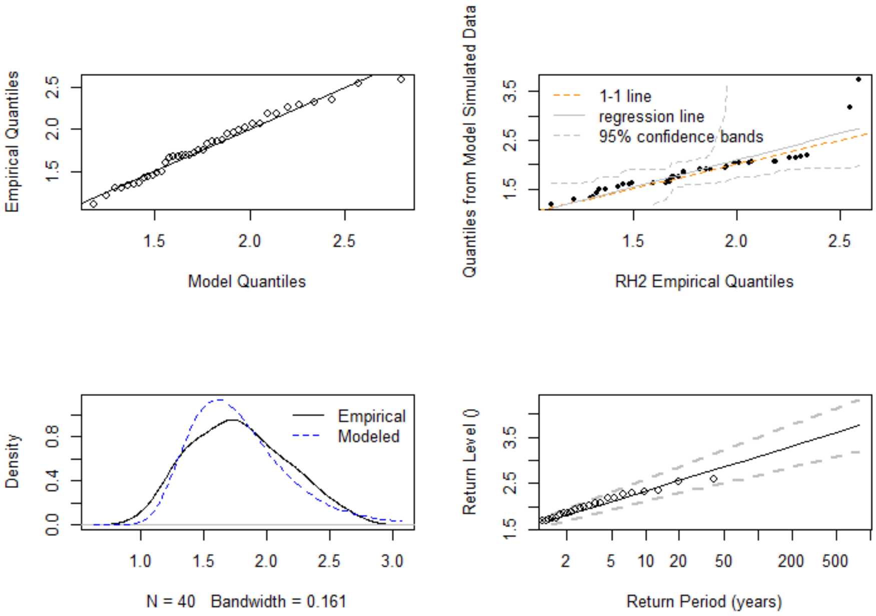

Figure 8 – Diagnostic plot for wave run-ups at location along the shore 2 with Gumbel distribution.

The quantile plot looks good, as in almost linear and fits the qq-line, where the empirical quantile plot seems slightly more spread out, but still decent. Density plot is good enough and return level over return period plot is almost linear on the plotted 50 years.

Over all, this point 2 is well fitted by the Gumbel distribution and therefore can be used later on.

Dune erosion

As said previously, the dune erosion data is showing various flaws, as many data points or whole sections of the coast have 0-valued arrays.

Coming back on the alternatives about the data treated in the parts (b.6) to (b.8), and using two years as a block size for running block maxima on the sets it shows identical results using either alternative ($b.7$), the maxima over the 3 main neighbouring spots Lats or ($b.8$), the maxima over the whole bay as the maxima over the vay and over those 3 spots are the same.

Figure 9 is a scatter plot for the dune erosion with 2 years maxima over the period 1976 − 2015 on the coast of Ängelholm. Data in centimeters.

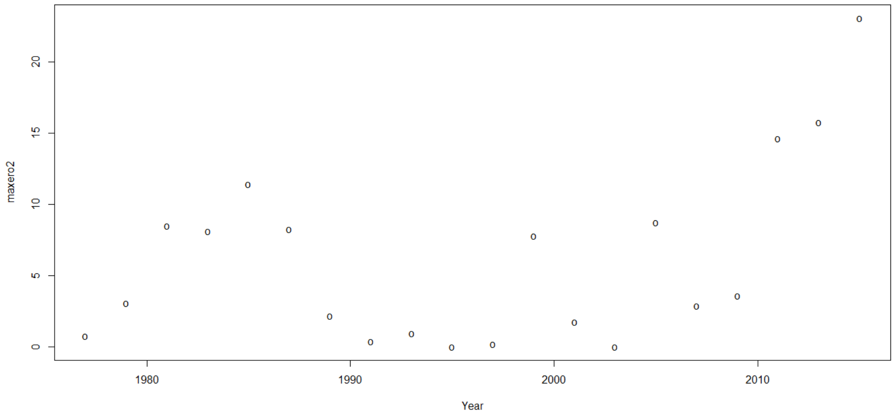

Figure 9 – Scatter plot for dune erosion in Ängelholm.

That was then used to create the following model using Gumbel distribution. Table 12 contains the parameter estimates and their 95% confidence intervals ; Table 13 contains 10−,100− and 200 return-levels with their 95% confidence intervals.

Table 5 – Parameter estimates for Gumbel distribution on 2-years maxima ; number in brackets are 95% confidence intervals for the parameter estimates.

location	scale
3.38 (1.42 ,5.33)	4.27(2.68,5.85)

Table 6 – Return levels for dune erosion with Gumbel distribution ; number in brackets are 95% confidence intervals for the return levels.

10-year return level	100-year return level	200-year return level
12.98 (8.43, 17.53)	23.00 (14.90, 31.10)	25.97(16.79, 35.15)

Figure 10 is a diagnostic plot for the Gumbel distribution with above parameters for maximum dune erosion on the bay of Ängelholm.

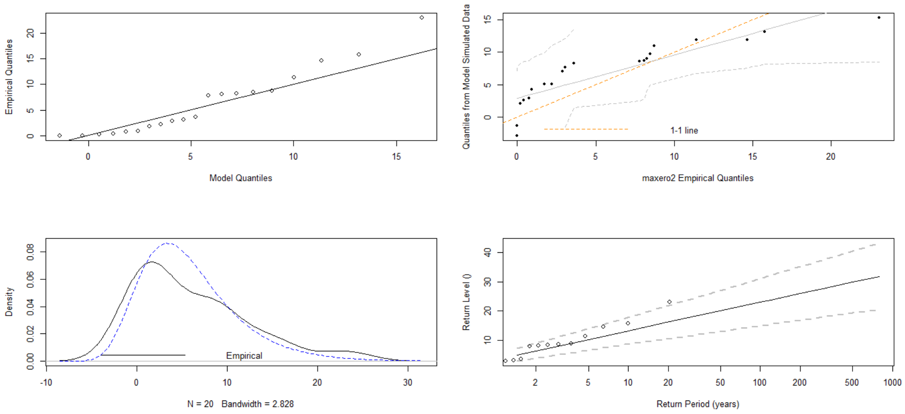

Figure 10 – Diagnostic plot for maximum dune erosion along the bay with Gumbel distribution.

The lack of data (20 points) also adds to the roughness of the plots, mainly for the model/empirical quantile plot which seems good. Return period plot and density against empirical distribution are good enough, where the empirical quantile against simulated model quantile is quite blurry.

Considering that 2 maxima out of 20 are still 0-valued and then decreasing the accuracy of the model, the goodness of fit for this model seems overall decent.

About wind speed and wind direction

The last part of that section will be dedicated to the analysis of wind components. Most of the problems coming from the data will be regarding to that section of this thesis. Here, how the wind component can be used to help creating a bivariate model for dune erosion using a wind component will be used. Data was in meter per second.

The data goes from January the 1^{st}, 1976 to December the 31^{st}, 2015 with a gap between December the 6^{th}, 2002 and January the 1^{st}, 2008 (so for 5 years, 24 days and 12 hours). Those are hourly measurements, hence it is a total of 44388 missing measurements. Because of the gap, the stationarity of the data was checked before analysing it. Using the R-package tseries and the R functions adf.test, pkss.test and PP.test, the set was deemed stationary enough (with a 95% confidence interval).

Figure 11 is a plot of Barkakra's wind speed against wind direction in degrees.

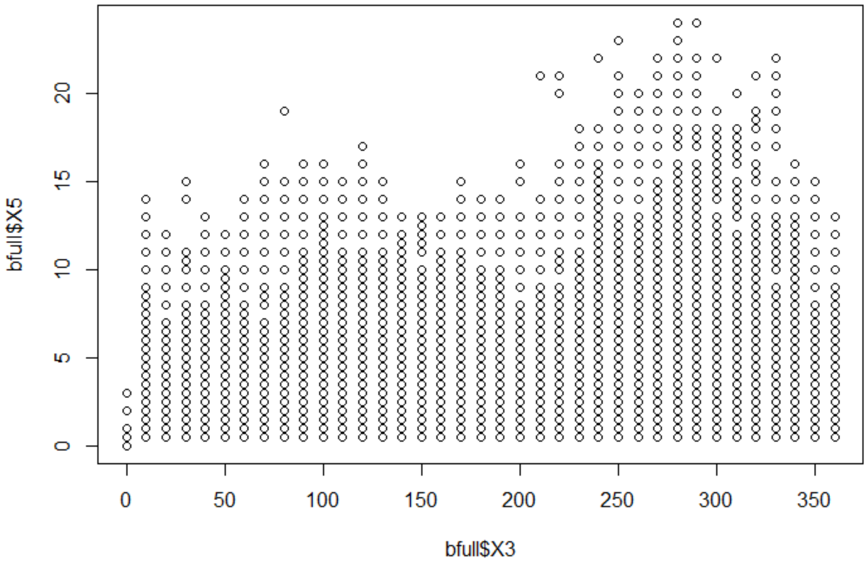

Figure 11 – Scatter plot of the wind direction against wind speed at Barkakra's station on the coast of Ängelholm.

From the Figure 11, it can be seen that the highest wind speeds come from the region between 210 and 330 which corresponds to inland wind blowing towards the sea. Some of the values from Barkakra's station were missing and were replaced by 0 to keep the same amount of data each year. They were not numerous so their absence shouldn't be significant as the orientation of the highest wind speeds is focused on the aforementioned region.

Before proceeding to analyse this data set, it was deemed important merge wind directions and wind speed altogether. See [**4**], as the same method was used here. Pertaining to the interest in extreme events, the aim was to create a set which shows the strongest wind as well as their direction. A wind blowing laterally to the dune should alter the dune less than one blowing perpendicularly. Although, the dune is shaped like a cave, so were used as a reference the directions mentioned above (210 to 330 degrees). By using a sinus transform of the direction, switched the sign to reward wind from that direction and then multiplying it to the wind speed. The obtained measurement is what will be called from now on, the oriented wind speed.

Proceeding now with extreme value analysis of this oriented wind speed, GPD model was used, see Table 7 for Scale and Shape parameter estimates.
Several thresholds were studied and having 10 as a threshold showed the best results.

Table 7 – Oriented wind speed with GPD ; number in brackets are 95% confidence intervals for the parameter estimates.

Scale	Shape
2.10 (2.02, 2.17)	-0.08 (-0.10 , -0.06)

Both parameter estimates are significant.
Table 8 shows the estimated return levels with their 95% confidence intervals.

Table 8 – Oriented wind speed with GPD ; number in brackets are 95% confidence intervals for the return levels.

10-year return level	100-year return level	200-year return level
17.61 (17.37, 17.86)	20.80 (20.29, 21.30)	21.65 (21.04, 22.25)

The spread of the confidence intervals seems to grow slowly and the results overall seem fair.

Figure 12 is a diagnostic plot for the oriented wind speed in Barkakra fitted with a GPD model.

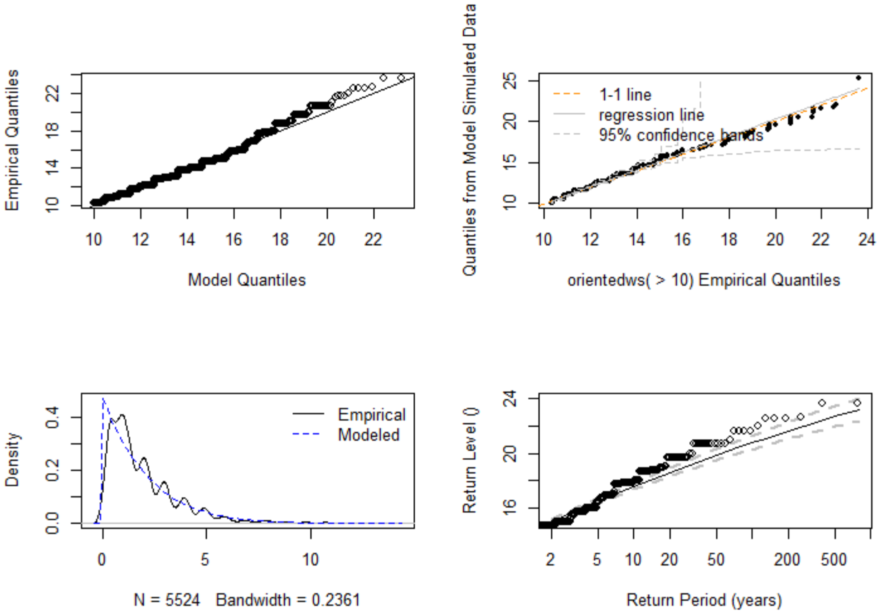

Figure 12 – Diagnostic plot for oriented wind speed at Barkakra station fitted with GPD.

Looking at the diagnostic plot, this model fits well the data. The quantile plot is almost linear and following the qq-line, same for the empirical one, density curve matches the theoretical one and the return period plot is good enough even though it seems to progress step by step.
One can conclude that a GPD model with 10 as threshold is a good fit for the oriented wind speed from Barkakra.

Although using a GPD model for the oriented wind data is not the best idea as a distribution from the same family as for the erosion, i.e. a GEV distribution, is used. Running a model for the oriented wind speed on a two-years bock maxima for GEV to match the erosion model and make the bivariate modeling doable.

Following is the analysis of the oriented wind speed for a two-years block maxima GEV model. Table 9 contains the parameter estimates along with their 95% confidence intervals.

Table 9 – Oriented wind speed with GEV ; number in brackets are 95% confidence intervals for the parameter estimates.

Parameters	Estimates
$\hat{\mu}$	19.25 (17.54,20.95)
$\hat{\sigma}$	3.59 (2.19,4.99)
$\hat{\xi}$	-0.62 (-0.94,-0.30)

All three parameter estimates are significant and a Gumbel model (shape $\xi = 0$) was rejected by maximul likelihood ratio test.

Figure 13 contains diagnostic plot from the previous model :

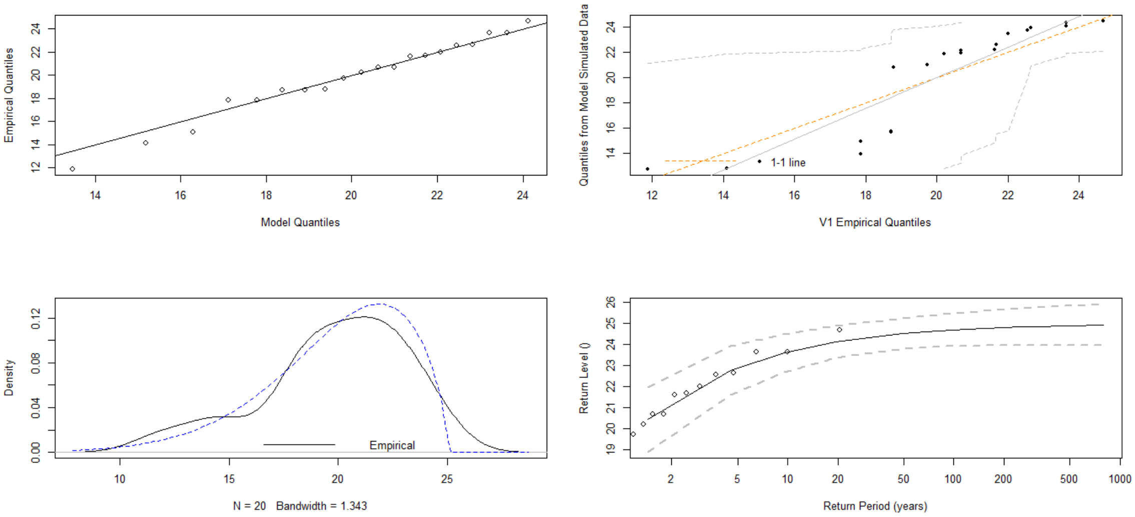

Figure 13 – Diagnostic plots for oriented wind speed on GEV model for two-years block maxima.

The empirical quantile against simulated data doesn't look good, but the others show a good fit to the data. One can say that the GEV model for two-years block maxima with parameters in Table 9 is a good model for our data.
With this model done, it is possible to proceed with a multivariate analysis, starting by trying to fit the nine parametric bivariate extreme value models to the data.

4.2 Multivariate analysis

As said previously, the main topic of this thesis is to model the effect of the wind on dune erosion. To do so, one wants to use a bivariate extreme value model. Because of the erosion data available (i.e. yearly maxima), only a bivariate block maxima can be fitted. Some information about the wind : the oriented wind speed is, as explained earlier, a combination of the speed and the direction of the wind, data from Barkakra station next to the shore of Ängelholm.

About the yearly erosion, and, as explained earlier : Most of the data points actually are 0's, making any analysis very difficult. Some spots Lats still have a lot of zeros or are completely zeros and therefore won't be used. They are mainly the points from 27 Lats to 59 Lats.
The amount of 0-valued points in the set was reduced by taking twice as large block maxima (i.e. from one year to two years block size). This is alternative (b.8). Even though half the original data size is rather low, the values now make more sense. Then, looking at what was left, and because of the interest in maximal values, it is required to choose data that is high enough for the analysis.

Now running all 9 models of bivariate (2-years sized) block maxima models, the following results were obtained.
Table 10 contains the parameter estimates for the 9 models and table 11 contains the dependence, asymmetry and alpha, beta parameter estimates for the same models.

Table 10 – Parameter estimates for location, scale, shape of both parts of the bivariate model, erosion and oriented wind speed, estimates respectively (1) and (2).

Models	location 1	scale 1	shape 1	location 2	scale 2	shape 2
"log"	1.87	2.69	0.97	19.04	3.56	-0.54
"alog"	1.85	2.63	0.94	19.18	3.55	-0.59
"hr"	1.79	2.56	0.96	19.07	3.56	-0.52
"neglog	1.82	2.61	0.95 "	19.06	3.56	-0.53
"aneglog"	1.86	2.62	0.93	19.20	3.56	-0.55
"bilog"	1.86	2.60	0.88	19.21	3.56	-0.55
"negbilog"	1.84	2.77	0.94	18.99	3.73	-0.55
"ct"	1.85	2.55	0.84	19.06	3.54	-0.54
"amix"	1.91	2.76	1.08	19.05	3.60	-0.58

The parameter estimates contained in table 10 are all close to each other, not depending on the chosen model. They all belong to the same kind of range with none really getting out of the group.

Table 11 – Parameter estimates for dependence, asymmetry and alpha, beta of both parts of the bivariate model, oriented wind speed and erosion.

Models	dep	asy-1	asy-2	alpha	beta
"log"	0.48				
"alog"	0.48	0.99	0.52		
"hr"	1.18				
"neglog"	0.73				
"aneglog"	0.86	0.99	0.63		
"bilog"				0.10	0.81
"negbilog"				2.49	0.10
"ct"				0.30	29.99
"amix"				1.18	-0.39

Table 12 – AIC and BIC for each of the 9 bivariate models.

Models	AIC	BIC
"log"	229.99	236.96
"alog"	231.41	240.37
"hr"	229.55	236.52
"neglog"	229.75	236.72
"aneglog"	233.20	242.16
"bilog"	226.92	234.88
"negbilog"	227.93	235.90
"ct"	229.08	237.05
"amix"	230.88	238.85

Looking at the lowest AIC and BIC, one can see that the "bilog" and "negbligog" models seem to be the ones most suited to fit our data sets. Although, as mentioned earlier, AIC and BIC are not goodness of fit tests, only criterions to differentiate models between them. A $p - value > 0.05$, or another goodness of fit test, is needed to be able to judge the significance of those models. As a consequence of using bivariate extreme value models, studying the dependence function of those models is a very good way to do so.

Figure 14 is a group plot of all 9 previous models' dependence functions. The fit is good enough if the empirical dependence functions matches the theoretical one.

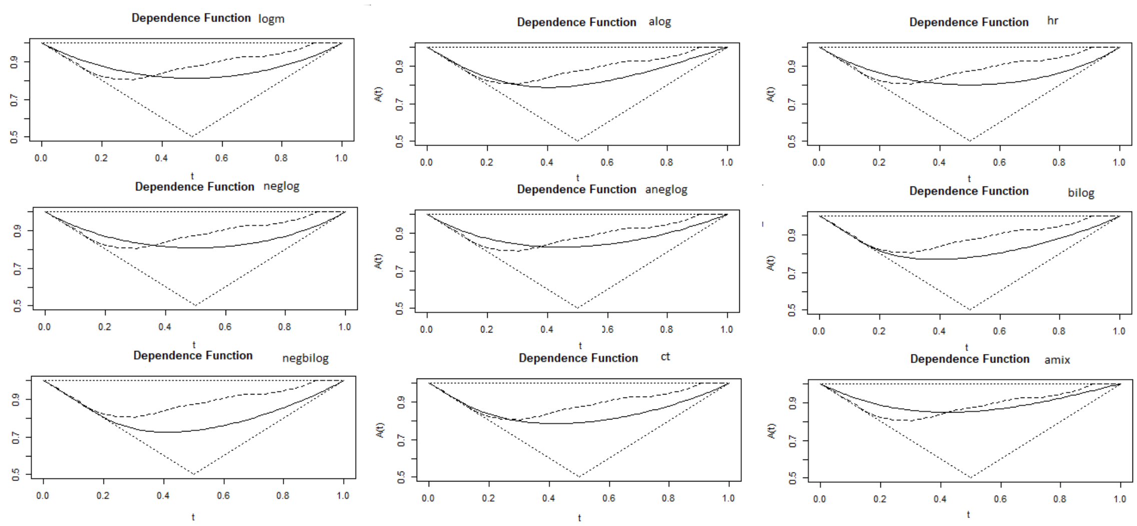

Figure 14 – Dependence functions for the 9 Bivariate Extreme value models listed above.

As one can see, from all 9 dependence functions plotted above in 14, none seems a good enough fit to the data. Those models are then deemed useless.
The "bilog model" can still be looked at. Figure 15 is a diagnostic plot for the the "bilog" model for the bivariate dune erosion/oriented wind. Being the best model (AIC and BIC) among the nine different ones, it is interesting to look at the diagnostic plot for this one.

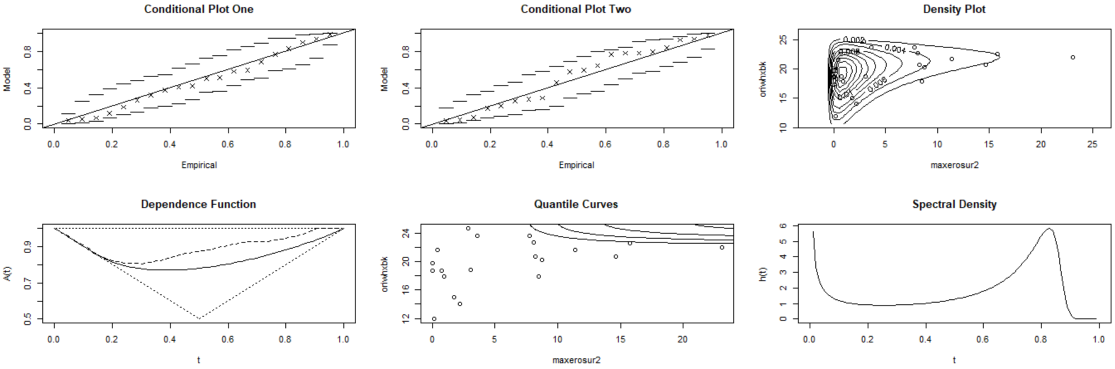

Figure 15 – Diagnostic plot for bilog bivariate extreme value model for dune erosion and oriented wind speed.

Once again, this model (bilog) is not statistically good enough. The previous plots are just interesting and so, worth mentioning, but since the classic parametric approach to bivariate extreme value isn't good enough for our data, a better model had to be created.

4.3 Copula estimation of bivariate extreme value model

As well as trying this new approach using copulas, the same data set as in the univariate analysis was used, i.e. a two-years block maxima for the oriented wind speed and on the whole coast erosion.

Using the marginal distributions for both erosion and oriented wind speed, respectively Gumbel and GEV distributions,the data set was transformed into a set of uniform $U[0,1]^2$ set. As explained in the theoretical part about Copulas, there are 5 families of extreme value copulas, Gumbel, Tawn, t, Galambos and Husler-Reiss. Each of them was tested using our uniform data set, using a goodness of fit test for extreme value copulas in R $gofEVCopula()$. For those which were deemed significant, an ad hoc copulas was created with the given parameters and their fit using AIC and BIC criterions was checked.
Following Table 13 contains the 5 EV Copula families along with the estimated parameters, $p - values, AIC$ and BIC.

Table 13 – EV Copula families with goodness of fit parameter estimates, p-values, and AIC, BIC fitted to data.

EV Copula Family	Parameter(s)	p-value	AIC	BIC
Gumbel	1.37	0.74	1.25	2.24
Galambos	0.64	0.79	1.16	2.16
Husler-Reiss	1.04	0.82	1.14	2.13
Tawn	0.70	0.88	1.55	2.54
t	(0.64, 4)	0.15	3.15	2 5.14

Note that the second parameter for the Extreme value t-distribution is the number of degrees of freedom. All 5 models have a $p - value > 0.05$ and are then significant. The smallest AIC and BIC come from the Husler-Reiss (hr) model.
A Husler-Reiss Copula with parameter $alpha = 0.718$ coming from the fit of the Husler-Reiss Copula to the uniform data set was used.

The following figures, Figures 16 to 18 show the density function, pdf and contour of the Husler-Reiss Extreme value Copula based model for our data set.

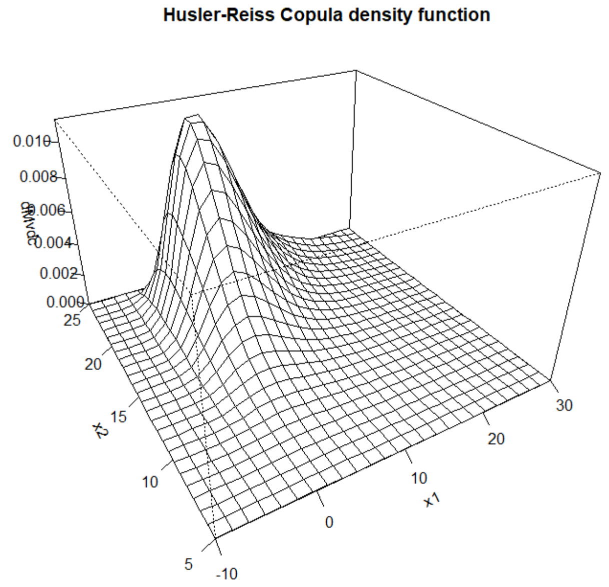

Figure 16 – Density function for the Husler-Reiss extreme value Copula based model for wind/erosion.

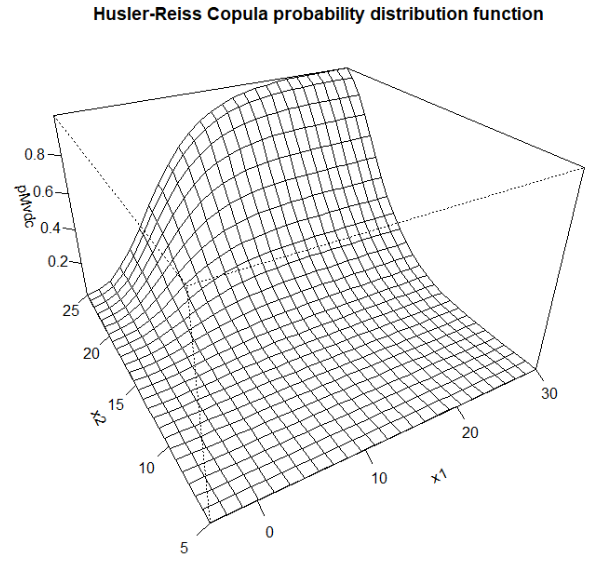

Figure 17 – Pdf for the Husler-Reiss extreme value Copula based model for wind/erosion.

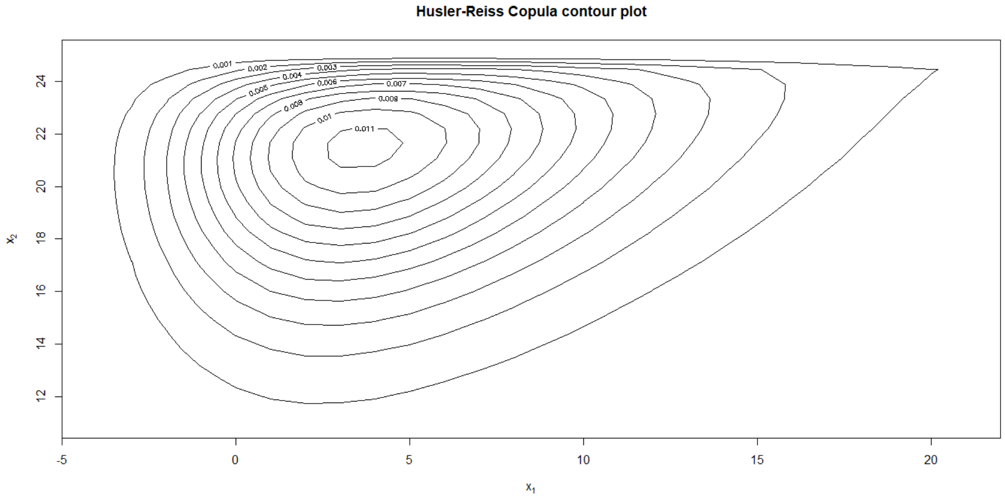

Figure 18 – Contour plot for the Husler-Reiss extreme value Copula based model for wind/erosion.

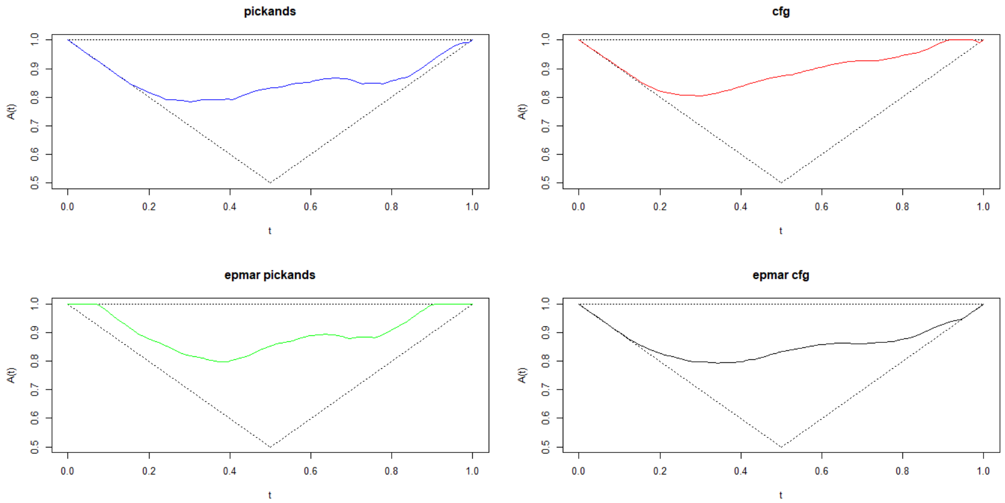

Figure 19 – Dependence functions for the Husler-Reiss extreme value Copula based model with pickands and cfg estimators, both with and without empirical transformation of the margins (epmar).

The resulting Husler-Reiss extreme value Copula based model is good enough ($p - value > 0.05$) but as there were only 20 observations used for the model and the dependence functions are not convex, we want to use smoothing splines to enhance the model.

Using the $R-$package *SimCop*, it is possible to create a non-parametric estimate of the dependence function will be added together with the previous Copula estimators of the dependence function.

In following Figure 20, the pink line is the smoothing spline estimate of the dependence function.

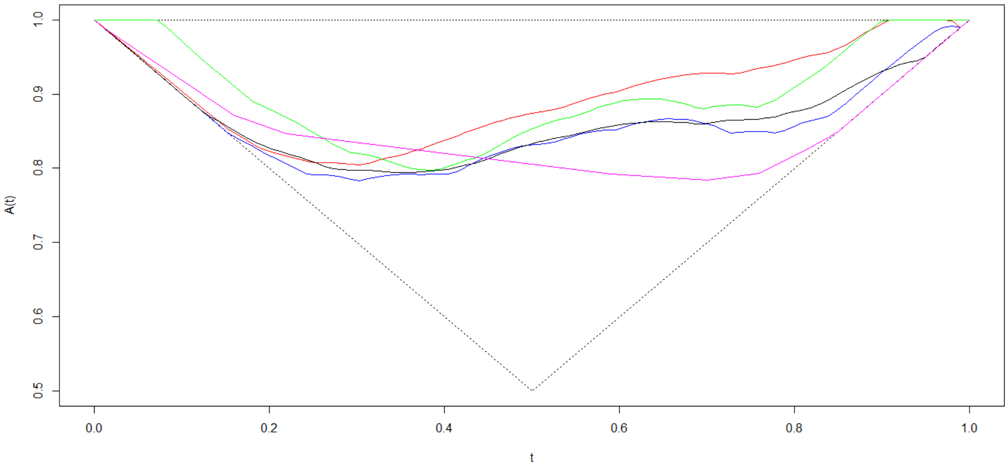

Figure 20 – Dependence functions for the Husler-Reiss extreme value Copula based model with pickands and cfg estimators, both with and without empirical transformation of the margins, epmar. Pink is the smoothing spline estimate and as is the previous plot, figure 19, blue is pickands estimator, red is cfg, green is epmar pickands and black is epmar cfg.

One can see that the latest estimate is a convex function, which is one of the three requirements of dependence functions that was wanted. This smoothing spline estimate is then the best one available for the already validated Husler-Reiss Copula based model.
Coming back to the lack of data, now proceeding with a parametric bootstrap to get better estimates of the parameters and return-levels of our bivariate wind/erosion model.
Using the non-parametric estimate of the dependence function, a smoothing spline can be fitted to create a new bivariate extreme value copula, approximate its density function and re-generate data.

Using a bootstrap with $B = 1000$ iterations to get as good estimates as possible.

Table 14 – Parameter estimates and 95% confidence interval for each of the margins using parametric bootstrap.

Parameters	CI lower bound	Estimate	CI upper bound
$\hat{\mu_1}$	1.25	3.38	5.10
$\hat{\sigma_1}$	2.14	4.27	5.99
$\hat{\xi_1}$	-2.12	0	1.73
$\hat{\mu_2}$	17.12	19.25	20.97
$\hat{\sigma_2}$	1.46	3.59	5.31
$\hat{\xi_1}$	-2.75	-0.62	1.10

Note that $\hat{\xi_2} = 0$ because the estimated marginal distribution is Gumbel, so with shape $\xi = 0$.The insignificance of the 95% confidence interval is not a problem. The same problem although, does happen with $\hat{\xi_2}$ which is not supposed to be null.

Apart from that, the results are decent and looking back at Tables 7, and 10 to 12, one can see that the Extreme value models have similar parameter estimates. Which, overall shows the consistency of the results.

The following plots, Figure (21) to (24) are contour plots for the bivariate return levels of oriented wind speed and erosion, our last and most wanted results. The results are in order showing $10-, 100-$ and 200-year return-levels. Remember that measurements were taken every two-year (two-years block maxima).

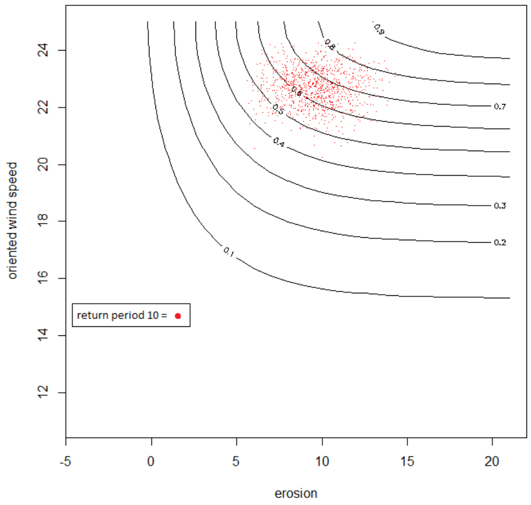

Figure 21 – Quantile curves and 10-year return levels by bootstrap estimation for wind/erosion model.

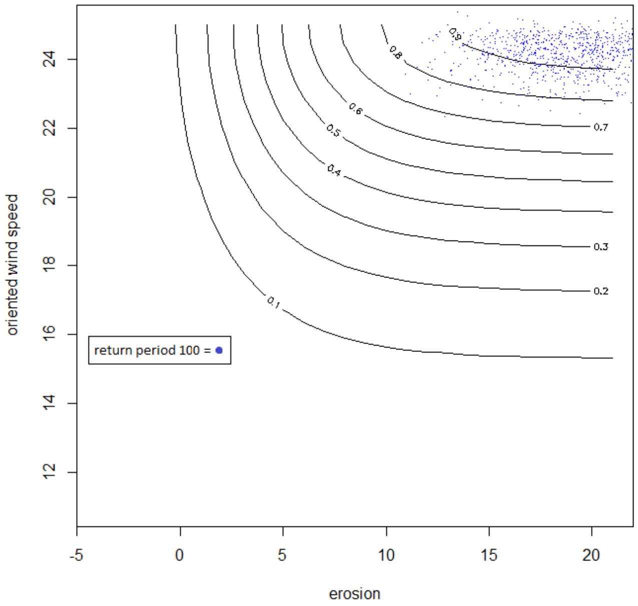

Figure 22 – Quantile curves and 100-year return levels by bootstrap estimation for wind/erosion model.

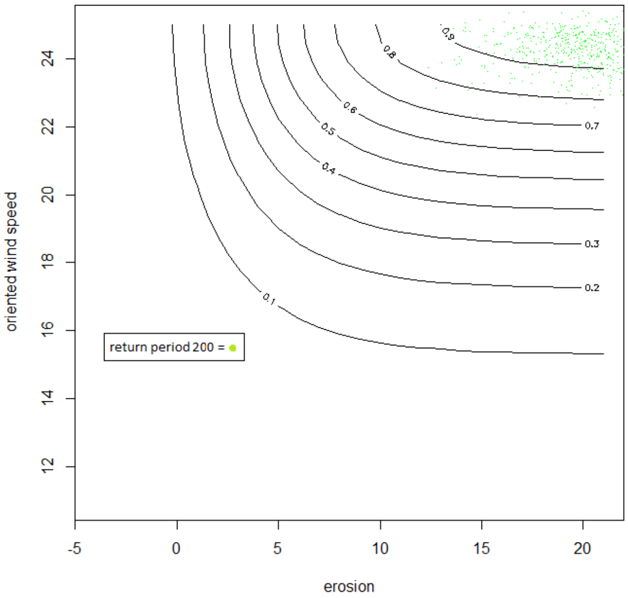

Figure 23 – Quantile curves and 200 year return levels by bootstrap estimation for wind/erosion model.

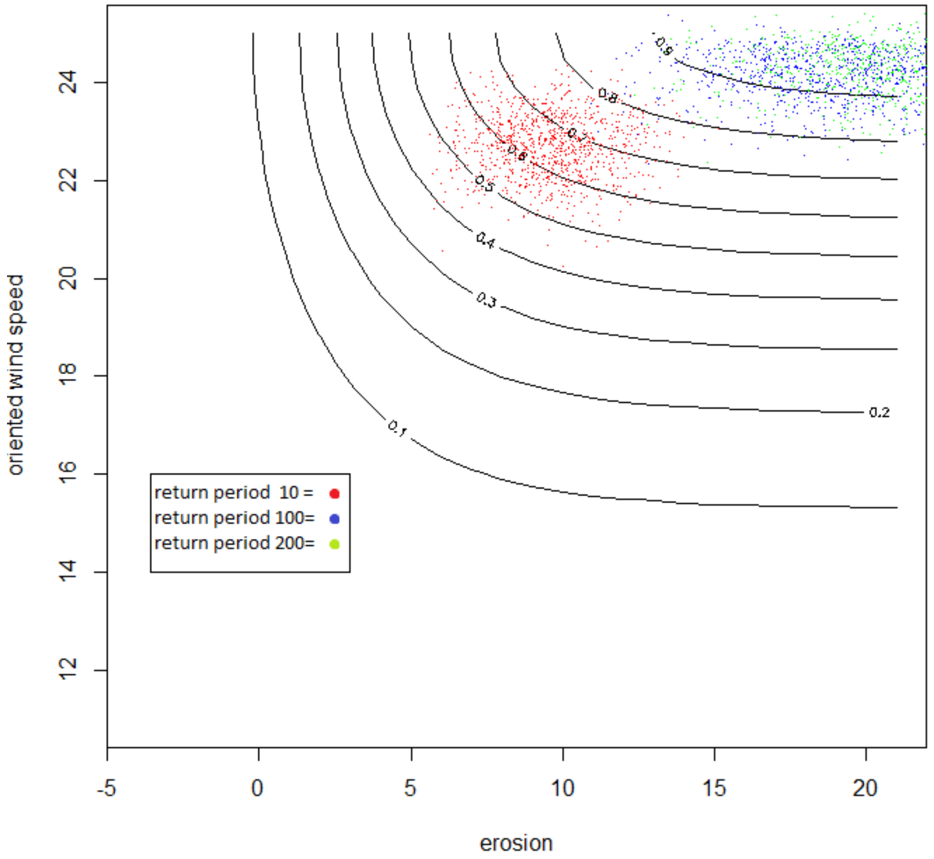

Figure 24 – Quantile curves and 10-, 100-, 200-year return levels by bootstrap estimation for wind/erosion model.Colors in the legend, respectively red, blue and green.

As explained right before the plots, having one measurement every two years makes the return period slightly different from usual as it is, in order for the $0.8, 0.98$ and 0.99 quantiles. One can see that the $10-$year return level, in Figure 21, appear around the 0.6 and 0.7 quantiles which is slightly lower that what was expected. Although, the following two plots, Figures 22 and 23 show bootstrap estimated return levels in around the ad hoc quantiles.

Table 15 contains the highest value of dune erosion and oriented wind speed in the future :

Table 15 – 10-,100- and 200-year return level for bivariate gev with Husler-Reiss copula based models. Number in brackets are their 95% confidence intervals from 1000 bootstrapped samples of data.

T	Erosion	Oriented wind speed
10	9.77 (6.68,13.07)	22.74 (19.65,26.03)
100	20.02 (16.93, 23.31)	24.50 (21.41,27.79)
200	23.0 (19.91,26.29)	24.68 (21.29,27.97)

These our are final results. As our model, the Husler-Reiss copula based GEV model for bivariate erosion and oriented windspeed was deemed significant by several goodness of fit tests, one can use those return levels as good enough predictions of what can happen in the corresponding future periods in the bay of Ängelholm, in Skåne.
Then, having that the maximum expected dune erosion there in the next 10 years is about 9.8 centimeters, 20 centimeters for the next 100 years and finally, about 23 centimeters in the next 200 years.

5 Conclusions, discussions and possible improvements.

The dune erosion level is an indicator of climate anthropy and can be used for flood risks prevision. In this report, several analysis were performed around the idea of such previsions but were mostly focused on a bivariate model between the dune erosion and wind speed. The dune erosion on the shore of Ängelholm's data, provided from 1976 to 2015, was on a yearly maxima format and thus, wind speed data needed to be cut down to this format as well. Using extreme-value theory, several models were fitted to the data to try and explain how each phenomena was behaving. Using block maxima on sea-level , wave runup, dune erosion and oriented wind speed with $1, 1, 2$ and 2 years blocks respectively. Afterwards, a model fitting both dune erosion and wind speed was created. Because of a lack of goodness of fit, copula theory was used to get a better model, using Husler-Reiss's bivariate extreme value copula. From the latter, several interesting return levels were calculated along with their confidence intervals using smoothing splines and boostrap method, giving an idea of how much erosion the dunes in Ängelholm would get in the future as well as the highest wind speeds the area would get.
The The analysis of the effect of wind on dune erosion in the bay of Ängelholm is now done with a significant model and return levels were calculated for three interesting periods of time.
The Husler-Reiss Copula-based Bivariate Extreme value model for oriented wind speed and dune erosion, helped by bootstrapping and smoothing splines means gave us estimates that can be used to avoid damages in the bay for the future times.

Referring to the diagram 2 with the dune, dune erosion, sea-water level and wave runup, one can now use the calculations made during the analysis of all those data sets.
Keep in mind that this part is not the main topic of this thesis but is here to show how to use almost directly the results from the bivariate model of the effect of wind on dune erosion. The calculations here will be kept simple. By using the 95% quantile of the extreme values distributions one can, by plain sums, get the following results :

118.91 centimeters for a 10-year period ;
227.84 centimeters for a 100-year period ;
and 241.861 centimeters for a 200-year period, which is the height of the coastal protections needed in the bay of Ängelholm in Skåne to avoid flooding in the future.

Remember that this part is just a simple calculation that is not made to be used straight away. The mere point of those is to show that the dune erosion, even though it represents a small part of those numbers is still important and thus should not be neglected.

Studying all of the 59 points located along the shore would not have given more insight or better results about the global effect of the wind on dune erosion on the coast of

Ängelholm, as it would mean running 59 analysis, which is a lot and seems like a rather inefficient way of analysing the data there.
Also, for the erosion, most of the points Lats have a nearly 0-valued data set, which is why a whole-coast-set was used instead of a single point. There is no use studying the points independently.

A way to get around the gap in Barkakra's data set would have been to reconstruct the missing values using more advanced techniques, such as, for example, a Kalman filter (alternative *b*.5). A Kalman filter is the exact solution of a state filtering problem for linear dynamic models ; see [**6**], *p*.289. Kalman filters usually needs data with assumptions about linearity and normality but according to [**6**] in appendix A, *p*.332, *it can be shown, using Hilbert space formalism that these predictions, and updates, are the optimal linear updates, even when the distributions are non-Gaussian.* Although, for this thesis, it was deemed unnecessary and rather tedious, considering that the reconstruction using Hallands Väderö A station's data was good enough. Reconstructing two points was worthless.
With more time, it would be interesting to study the following points or topics :

— gathering more data on dune erosion and dune height to make more accurate models where a reconstruction would then be useful ;
— looking into the SMHI models for their unverified data, it would then be possible to use only the verified points or the whole set with more insight of their meaning ;
— trying to model other climatic phenomenon with the erosion to see if there have more impact on it than the wind.

Bibliographie

I would like to thank the authors of the following books, articles, thesis, packages and documents which helped me along the way in my own thesis :

[**1**] S. Coles (2001), "An introduction to statistical modeling of extreme values" ;

[**2**] R.B. Nelsen (2006) ; "An introduction to copulas"

[**3**] C. Hallin (2019), "Long-term beach and dune evolution, Development and application of the CS-model" ;

[**4**] K. Persson (2017), "On risk analysis of extreme sea levels in Falsterbo peninsula" ;

[**5**] C. Maia (2010), "Multivariate Empirical Cumulative Distribution Functions", code and package ;

[**6**] H. Madsen, E. Lindstrm, J.N Nielsen (2015), "Statistics for finance" ;

[**7**] A. Jakobsson (2016), "An introduction to time series modeling" ;

[**8**] M. Sandsten, G. Lindgren, H. Rootzn (2014), "Stationary stochastic processes for scientists and engineers" ;