Interreg
North Sea Region
FAIR
European Regional Development Fund
EUROPEAN UNION

Miljø- og
Fødevareministeriet
Kystdirektoratet

Joint Probability method report

2019

Publisher:
The Danish Coastal Authority
Højbovej 1
7620 Lemvig
www.kyst.dk

Publishing date:
April 2019

Authors:
Charlotte Ditlevsen
Ulf Radu Ciocan

Illustrations and layout:
Charlotte Ditlevsen
Ulf Radu Ciocan
Marta Merino

Front page graphics
Ragnhild Kørschen

Project Responsible:
Ane Høiberg Nielsen

Project Manager:
Ulf Radu Ciocan

Keywords:
Statistik, oversvømmelse, hav, vandløb, risiko.

Contents

Summary

Presented here is an introduction to joint probability statistics and the results for one pilot site in Ribe, Denmark.

The joint probability method is performed on sea- and stream-water level data. The analysis includes data from 5 water level measure stations in Ribe å (4) and Hjortvad å (1).

Through the development and evolution of the method, several thoughts and ideas were processed and most are presented within the discussion section.

The results are to be used in flooding models at a later time.

The method has its flaws, as sea data is given a higher weighting compared to the stream data, but not much can be said without further analysis, especially on the effect of stream data as first variable.

It is advised to further analyze the method and results and look into other variables, in regard to flooding sources, especially, finding the effect of the method's weighting of the first variable.

1. Introduction

Many coastal cities are affected by coastline recession and storm erosion from high water levels and waves. Over time, in many places, this has led to the construction of coastal defenses such as sea walls and revetments to prevent this erosion. Additionally, coastal flooding from ocean extremes is a major concern in many low lying coastal areas and where flood prevention e.g. in the form of dikes and interim measures serve to reduce negative impacts of flooding for the society at large.

In Flood Risk Management (FRM) and Flood Risk Reduction (FRR) one initial approach is to analyze the hazard and the potential source and cause of flooding. The source of flooding can be the sea, precipitation, rivers and streams etc. from meteorological extremes, and the cause may relate e.g. to the failure of dikes. In cases where more than one source can lead to flooding, it is valuable to gain information about potential coincidences in extremes between the sources and to establish to which extent a conjunction of extremes may lead to more severe floods in terms of lateral extention,flooding depths etc. The joint probability of simultaneous flooding from streams and the ocean at a given location is the main focus of the present work, which seeks to describe the likelihood of a scenario involving two sources, which, when combined constitute an even greater hazard than they would individually. Naturally, the importance of such work can be crucial in the determination of risk reduction efforts and for coastal towns with intersecting rivers and/or streams which make them more prone to floods.

Based on Danish cases and hydrological conditions, the joint probabilities between different water levels in the sea and corresponding water levels in streams over time are investigated. This ideally represents realistic flooding hazards (or threats) and the results may enter hydrodynamic flood modeling for further exposure and risk mapping.

The sources that are taken into account are: medium-high to extreme water level events in 1) the sea and the coinciding 2) stream data. Different combinations of water levels in both sources are evaluated as to their statistical return periods and return water levels. Also, potential changes in water levels as a result of climate change are taken into consideration.

To find the potential extent of a flood, a model is applied to visually/graphically demonstrate what occurs in a certain area when sea and stream water levels reach a critical level or threshold. The model parameter inputs are: return value, return period,the total mass of water entering the system and the duration of the flooding, and the model analysis provide most of these data. All statistics presented are only as robust as the data from which they originally derive and this must be kept in mind when interpreting the results.

The report is delivered as part of the 'Flood infrastructure Asset Management and Investment in Renovation, Adaption, Optimization and Maintenance' (FAIR) Project, which is a flood risk reduction project co-funded by the EU Interreg North Sea Region Program 2014-2020.

1.1 Prerequisites

This report presents advanced statistical methods and therefore requires some understanding of and/or interest in statistic marginal distribution functions, as well as dependence structures between sources and return periods, in order to fully appreciate discussions and results. For further information about the applied statistics, please consult the literature referred to in the text.

2. Statistical understanding

When applying statistics to a problem, the outcome is typically one or more of the following: organization, analysis, and presentation of data (graphically) and/or their interpretation. Statistical analyses are often applied to identify tendencies and trends in the future. In other words, if we know all about the tendencies in a process, and we know how often it happens, we have the opportunity to ”predict” the future as a result of expected tendencies. This report has a particular focus on “test of the relationship”, the relationship being that of sea water levels in relation to the water level of streams.

2.1 “Good” statistics

Good statistics are a combination of 1) good data and 2) robust statistics (statistics with good performance for data plotted from a wide range of probability distributions).

Outliers are data points with values far from the mean of the total data sample or even contradictory to the tendency in the other data points, so that they appear to “stand alone”. It is important to handle outliers and the way they are handled is decided by the conductor of the study. If remaining in the sample, outliers should be actual measurements of value; these points could obscure the fit of some marginal distributions. Some statistics are highly affected by outliers and in observing outliers, particularly in water level data, one needs to apply probability distributions that are insensitive to relatively small deviations from the assumptions associated with the dataset. The 2-parameters distribution functions (e.g. LogNormal) are typically less sensitive compared to the 3-parameters (e.g. GEV).A more thorough description of data and the choice of distribution etc. is provided in section 7.3.4 below.

2.2 Data

Half of what it takes to make “good” statistics is good data and, as such, it is important to evaluate data. Taking ocean water levels as an example, the following is required, or at the very least, highly recommended:

1. Where is data measured?
   A location description is highly important since the geophysical circumstances often affect the measurements.

   Example:

   Depending on the location of the instrumentation, e.g. a tide gauge station, winds have an effect on the water in its immediate surrounding. For instance, at the town of Ribe on the Danish Wadden Sea coast, onshore westerly winds lead to a raised water level, as water is pushed towards the shore (wind set-up). The tide gauge station at Ribe is placed adjacent to the sluice in the dike, and during high water levels in the sea the sluice gate is closed. This means that sea water is pushed against the sluice walls and to some extent accumulates near the sluice, affecting the actual water level registration to indicate a higher than actual mean sea level (MSL).

2. When and how often is data measured?
   Not all data follow the same measurement intensity (some water levels are measured every 10 min others once a day). The higher the intensity, the more precise is the measurement (given that the instrument is measuring correctly).

3. How long is the dataset?
   The general rule is: the longer the dataset, the better the statistics. This, of course, depends also on the quality of data etc. For most statistics related to extreme water levels Kunz, Flügge, and Franzius (2007) found that a calculated return period should be no more than two to three times the actual

dataset length. This means that 20 years of observation may approximately represent what data will look like 40-60 years later. Most "natural" parameters (precipitation, ocean water level, temperature, pressure etc.) are affected by climate change tendencies and these parameters typically need a progression factor to account for changes over time.

4. How many missing data points or outliers ?
   If a data series contains too many gaps, this leads to less representative results due to potentially missing extremes. Some outliers are measurement errors and should be removed, but not all outliers are errors, however. It is thus very important to understand which data are real and which are not.

5. In which datum is data noted?
   When working with water level data it is important to know to which datum data are referenced. Denmark currently uses the national height reference system Danish Vertical Reference 1990 (DVR90). The DVR90 datum zero corresponds to mean sea level in Denmark in 1990. The common national reference allows all water level data to be compared in terms of water level.

6. Data detrending
   Due to past and ongoing climatic changes, the mean water level is not constant. Additionally, post-glacial rebound (glacio-isostatic adjustment) of the Earth's crust is experienced as vertical land motion and needs to be accounted for in tide gauge records. The detrending of water level data serves to reference all data to the same zero. Detrending procedures are further explained in section 4.2.

2.3 Introduction to Marginal Distributions – Univariate statistics

The univariate marginal distributions give the probabilities of various values of the variables in the individual datasets, without reference to the values of the joint probability. In other words, the probability, return value and return period are found by marginal distribution.

2.4 Introduction to Joint Probability – Bivariate Statistics – Copula

Joint probability is the probability of two or more events occurring within the same time interval. The copula is a multivariate probability distribution that describes the dependency structure between random variables independently from their marginal probability distributions. It basically links up the univariate marginal distributions through their dependency structure.

A bivariate (joint probability) distribution describes or defines the simultaneous behavior of a bivariate sample. Whether data is A. discrete or B. continuous, the distribution can be described with; A. a joint probability mass function or B. a joint probability density function.

There are many copula families, however, the most commonly used for hydrological data is the Archimedean copula family, which includes, amongst others, three parameter functions, the Clayton-, Frank- and Joe distributions or the two-parameter copula functions (BB1 or BB6).

3. Location and Data

The following sections describe the physical environment of the case study area in Denmark and data used in the presented work.

3.1 Ribe, Location Description

The pilot site of Ribe is located on the Danish Wadden Sea coast. The town of Ribe is situated approximately 7 kilometers inland from the coast and is known to be oldest town in Denmark. A present and future challenge of the low-lying Ribe area is to rebuild or renovate the dikes to maintain the required safety against flooding of the hinterland and the town.

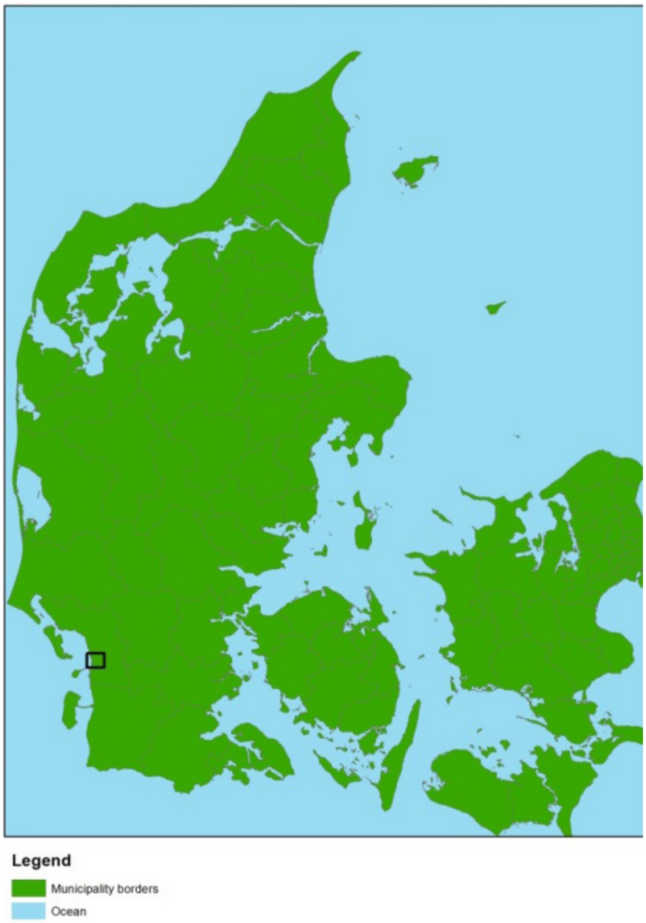


Figure 1 Location of pilot area

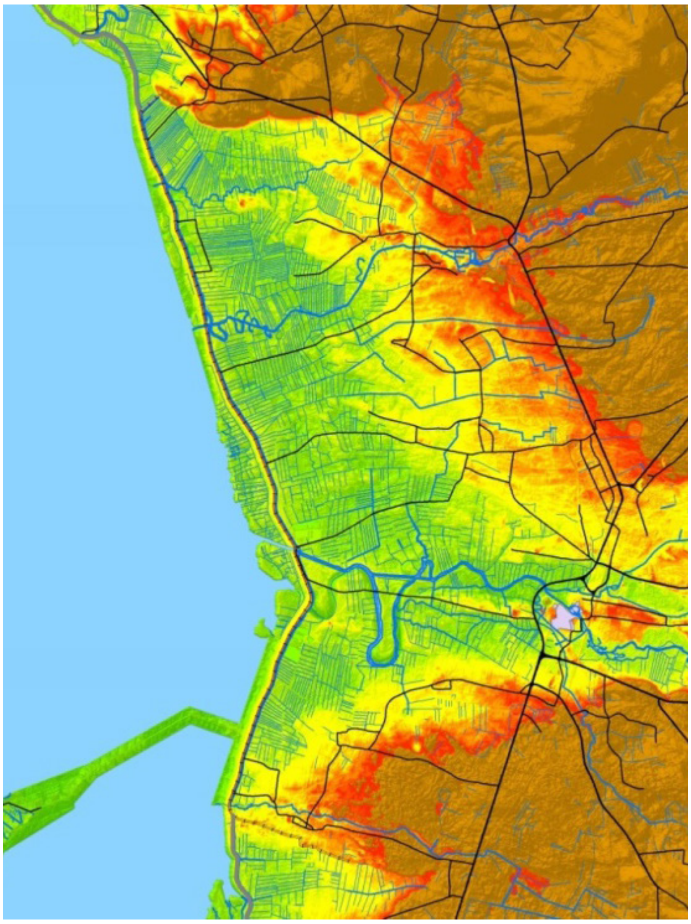


Figure 2 Height model with dikes, water courses and main roads

The landscape between the North Sea/Wadden Sea and Ribe is characterized by being a polder with marshes and moors. During the second to last ice age the moorland was formed. Later, sea level rise facilitated a change from moorland to marsh. The furthest inland extension of the marsh is 5 km at the stream, Ribe Å.

As shown on figure 2, sea dikes extend along the Wadden Sea coast as a flood protection barrier, only interrupted by sluices. Two streams pass through the pilot site; Kongeåen and the abovementioned Ribe Å. The area around Ribe is a large low lying (1.0-2.0 m DVR90) flood plain containing mostly drained farmland. The total area of the flood plain is 97 km2. At the inland margin of the flood plain, the town of

Ribe is located. The old town center is elevated above 4.0 m DVR90, but the town has developed and also includes residential and industrial areas located as low as 2.0 m DVR90.

Several flood protection structures are found in the pilot site of Ribe, including dikes, sluices and pumps. The flood risk in the Ribe flood plain is two-fold. One risk stems from the sea, consisting of potential dike breaches or overtopping during storm surges. The other risk is flooding fromthe local streams. The main sources of flooding from watercourses are two streams, Ribe Å and Kongeåen, and two smaller creeks which have their outlet in the Wadden Sea through the dike either by lock or sluice. Ribe Å flows through the old town centre of Ribe and through the drained farmland and finally passes the dike through the lock, Kammerslusen, into the Wadden Sea. Flooding from groundwater and precipitation is not considered, and only flooding from Ribe Å is included in the presented study.

During storm surges, extreme high water levels in this part of the Wadden Sea may reach 5.0 meters DVR90 (which corresponds to a 100-year return period) and exceed the terrain height of most of the Ribe flood plain.

During a storm on the 3rd of December 1999 water levels at Kammerslusen reached 5.10 m before the tide gauge broke down. The storm surge peak time coincided with astronomical low tide and no dike breaches occurred.

The tidal amplitude is 1.5-2.0 m in the Danish Wadden Sea. Had the peak of the surge coincided with high tide during the 1999 event, the expected water level would have exceeded 6.0 m. With a dike crest height of 6.88 m, wave overtopping would have been severe and the dike may have breached.

During storm surges, the lock and sluices in the dike line are closed and inland water levels rise due to the discharge in the streams. This indicates that the probability of flooding from the streams is at its highest during storm surges, or, in periods with elevated ocean water levels.

When the Ribe Å cannot discharge into the Wadden Sea this leads to flooding of the low lying area around the town of Ribe. This has happened on several occasions over the past decades.

The diked marsh is protected from the otherwise frequent flooding of sea water, and some of the marshlands are sufficiently drained to utilize for cultivation of grain. However, the typical terrain in the marsh is only about 1.5 m DVR90 in this low relief landscape. The meadows around the river valley are therefore too wet to cultivate and are used for hay and grazing. It is not unusual to see the meadows being partly or entirely flooded after high stream discharge events, heavy rain and/or if the sluice has been kept closed over a prolonged period due to high ocean water levels.

The river valley is dominated by Ribe Å that meanders towards the Wadden Sea. The stream is fed by creeks and brooks and in several places by smaller streams and canals as well. Just west of Ribe, the stream has been hedged into artificial canals to facilitate boating . Within the town of Ribe, the stream has also been impacted by human interference. The oldest part of the town is built on a dam. Canals and watermills are established within the town perimeter. At high runoff from Kongeåen, excess water flows through a canal to Ribe Å causing further pressure on the system. Installing pumps at Kongeåen has been considered, but so far financial resources have not been available for such measures.

3.2 Data Foundation

Figure 3 shows a map of the hydrological stations included in the calculations of the joint probability. The stations are mainly in Ribe Å (4 stations) and one station in Hjortvad Å.

The tide gauge data series from Ribe is approximately 98 years long (measurements started 5th of December 1919), but only the overlapping dates between sea and stream are used in any calculation of joint probability.

For the stream data the stations were selected based on their location in or in the vicinity of Ribe.

- The first criterion is a minimum time series length of five years, or rather, five hydrological years. The hydrological year is defined as the period from the 1st of October to the 1st of April which is the season that has the higest frequency of storm surges and heavy precipitation .

- The second criterion is a minimum of 75 % of a hydrological year represented in the data series. If a hydrological year lacks more than 25 % of the data, that hydro year is discarded.

- The third criterion is that all outliers are removed from the data series. It is carefully considered whether a data point is an outlier caused by an error of measurement or is, indeed, a measurement from an extreme event.

See table 1 below for information on each data series.

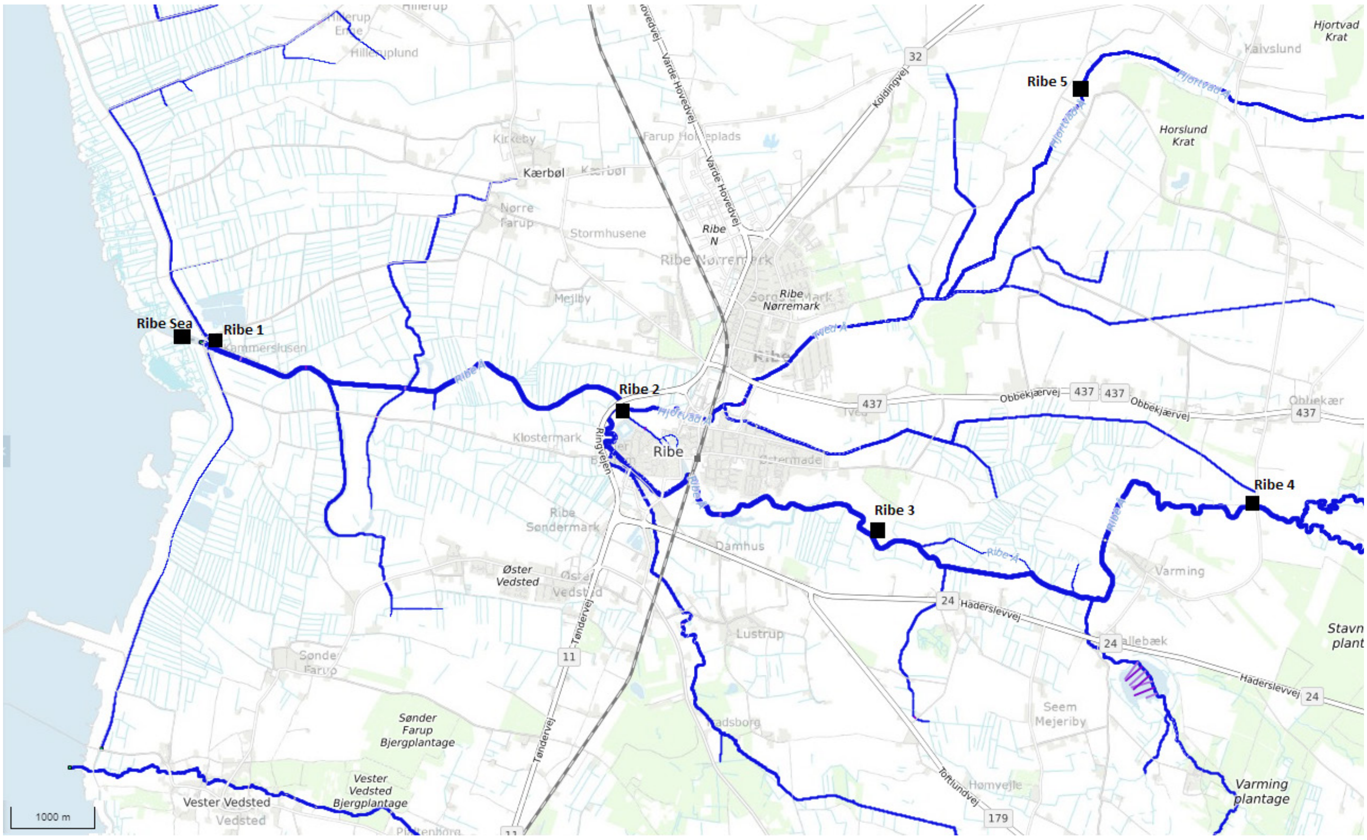


Figure 3 Map of the test site showing all stations marked on to it. It shows that 4 out of 5 stations are located in Ribe Å and 1 in Hjortvad Å.

Ribe stations	Data series length	Corodinates (UTM Zone 32N)	
Ribe 1	01.01.2004 - 08.12.2017	479520	6132645
Ribe 2	01.01.2004 - 08.12.2017	484712	6131689
Ribe 3	01.01.2004 - 04.01.2017	489747	6129936
Ribe 4	01.01.2004 - 24.02.2017	4822477	6130910

Table 1 The stream water level (gauges) , positions and respective time series

3.3 Assumptions and Uncertainties Associated with the Input Data

These are assumptions made for the data series:

1. The measurements are performed correctly and have been corrected, if needed, before delivery and use.

2. The data series are representative and reflects the long-term behavior within the catchment area. They contain a sufficient number of all types of events (low as well as high) and extreme events, in particular.

These are the uncertainties:

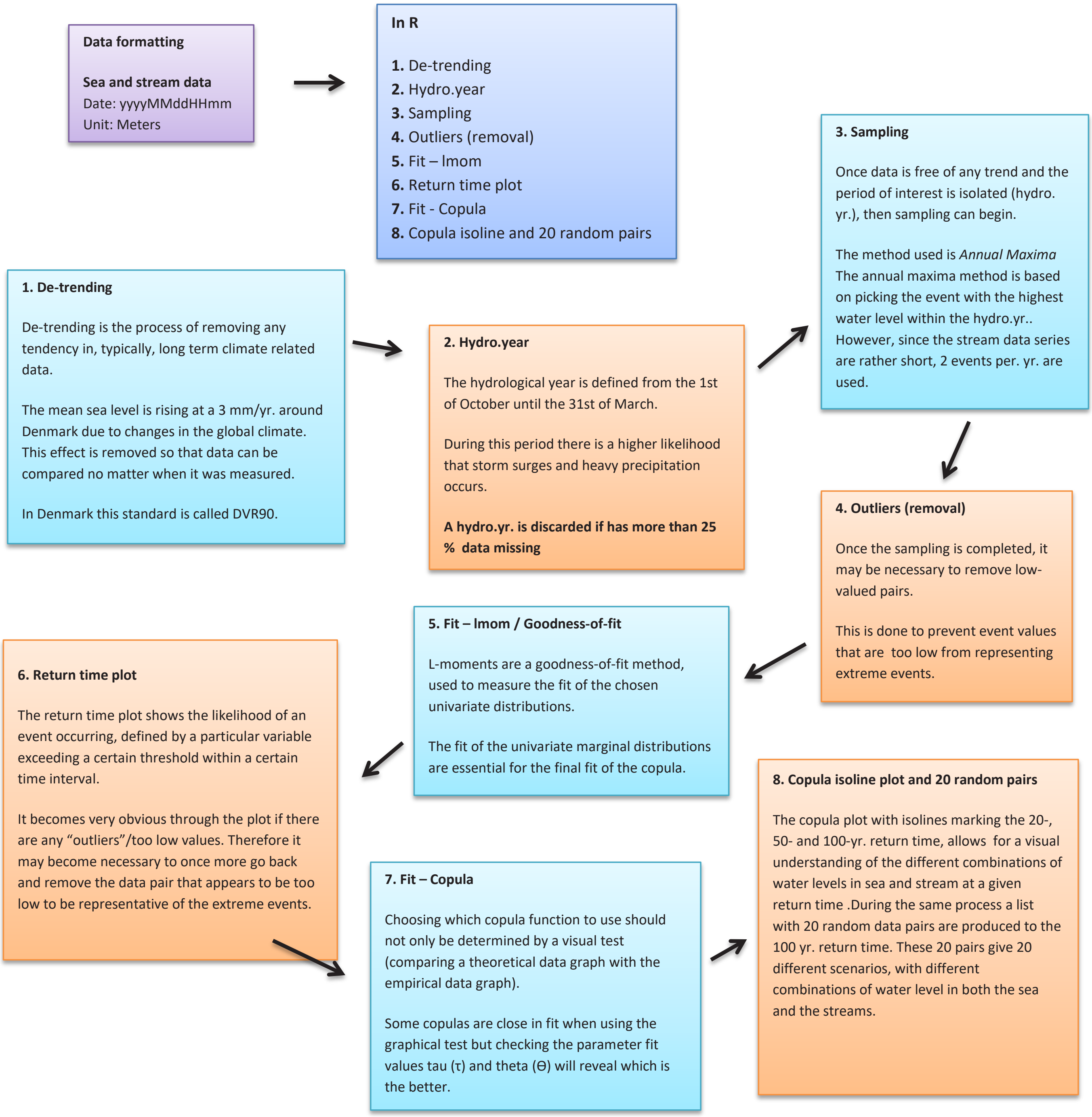
1. Some of the shorter series on stream data do not necessarily contain any extreme events

   The Danish Coastal Authority (DCA) produces extreme statistics on sea data every 5th year. The latest version is from 2017 (HVS17) and contains values for a 20, 50 and 100 year expected return period. However, the statistics in HVS17 are based on much longer data series and are likely to be more representative than the shorter data series used in these calculations. The reason for the shorter sea data series is that only approximately 20 % of the newest sea data series overlap with the existing stream data. Fewer extremes in the actual sea data series analyzed yields a lower and misleading extreme return value. In other words there is a risk that the return values might not be representative.

4. Method – Statistical Analysis

All probability calculations are performed in R. R is a software environment and language used for statistical computing and graphics. It is an Open Source program which provides extensive statistical and graphical techniques (r-project.org).

The complete joint probability process is shown in figure 4 and described below.



Data formatting

Sea and stream data
Date: yyyyMMddHHmm
Unit: Meters

In R

1. De-trending
2. Hydro.year
3. Sampling
4. Outliers (removal)
5. Fit – lmom
6. Return time plot
7. Fit - Copula
8. Copula isoline and 20 random pairs

1. De-trending

De-trending is the process of removing any tendency in, typically, long term climate related data.

The mean sea level is rising at a 3 mm/yr. around Denmark due to changes in the global climate. This effect is removed so that data can be compared no matter when it was measured.

In Denmark this standard is called DVR90.

2. Hydro.year

The hydrological year is defined from the 1st of October until the 31st of March.

During this period there is a higher likelihood that storm surges and heavy precipitation occurs.

A hydro.yr. is discarded if has more than 25 % data missing

3. Sampling

Once data is free of any trend and the period of interest is isolated (hydro. yr.), then sampling can begin.

The method used is *Annual Maxima*
The annual maxima method is based on picking the event with the highest water level within the hydro.yr.. However, since the stream data series are rather short, 2 events per. yr. are used.

4. Outliers (removal)

Once the sampling is completed, it may be necessary to remove low-valued pairs.

This is done to prevent event values that are too low from representing extreme events.

5. Fit – lmom / Goodness-of-fit

L-moments are a goodness-of-fit method, used to measure the fit of the chosen univariate distributions.

The fit of the univariate marginal distributions are essential for the final fit of the copula.

6. Return time plot

The return time plot shows the likelihood of an event occurring, defined by a particular variable exceeding a certain threshold within a certain time interval.

It becomes very obvious through the plot if there are any "outliers"/too low values. Therefore it may become necessary to once more go back and remove the data pair that appears to be too low to be representative of the extreme events.

7. Fit – Copula

Choosing which copula function to use should not only be determined by a visual test (comparing a theoretical data graph with the empirical data graph).

Some copulas are close in fit when using the graphical test but checking the parameter fit values tau (τ) and theta (ϴ) will reveal which is the better.

8. Copula isoline plot and 20 random pairs

The copula plot with isolines marking the 20-, 50- and 100-yr. return time, allows for a visual understanding of the different combinations of water levels in sea and stream at a given return time .During the same process a list with 20 random data pairs are produced to the 100 yr. return time. These 20 pairs give 20 different scenarios, with different combinations of water level in both the sea and the streams.

Figure 4 The process of the joint probability calculations.

4.1 Data preparation

Data is changed into a uniform format (date: yyyyMMddHHmm / year month day hour minute) with an hourly measurement (for some stations with a higher measurement frequency, an average for the full hour is found and used). Water level measurements are shown in meters. Data should already be free from errors, systematic errors, inaccuracies and outliers but, if located, further errors are removed.

4.2 Detrending

The detrending process is necessary due to changes in MSL over time. A constant rate of change of 3 mm/year (Dangendorf et al. 2017) along the Danish coastline is used, and its effect is removed in the data. At Ribe the vertical land motion is close to zero and is disregarded (Knudsen, Khan, Engsager, & Sorensen, C. (2016)).

4.3 Sampling

These are the basic requirements of a random sample:

a) All samples must consist of homogeneous data

b) All variables must be identically distributed

c) The events must be independent from each other (there is a minimum of 72 hours between events to ensure that the latter event is not influenced by the former

d) Stream data is found within +/-12 hours of the stream event (see figure 6 below)

e) No registration errors in the sample

f) The sample should reflect extreme events in its population

g) Stationary dataset

Annual Maxima

For choosing the samples, the Annual Maxima (AM) (or Block Maxima) method is used with two events per year. This method uses the largest event in each year as a sample. Because the data series are short it can be difficult to get enough data points to make a representative distribution function fit and therefore it can be an advantage to take the two largest events in each year. However, if there is a large difference in the two events and only one is representative for the sample, then the other data pair, or the outlier, can be discarded. For more information on block maxima and extreme data, please refer to Maximum likelihood estimators for the extreme value index based on the block maxima method - Dombry, 2013.

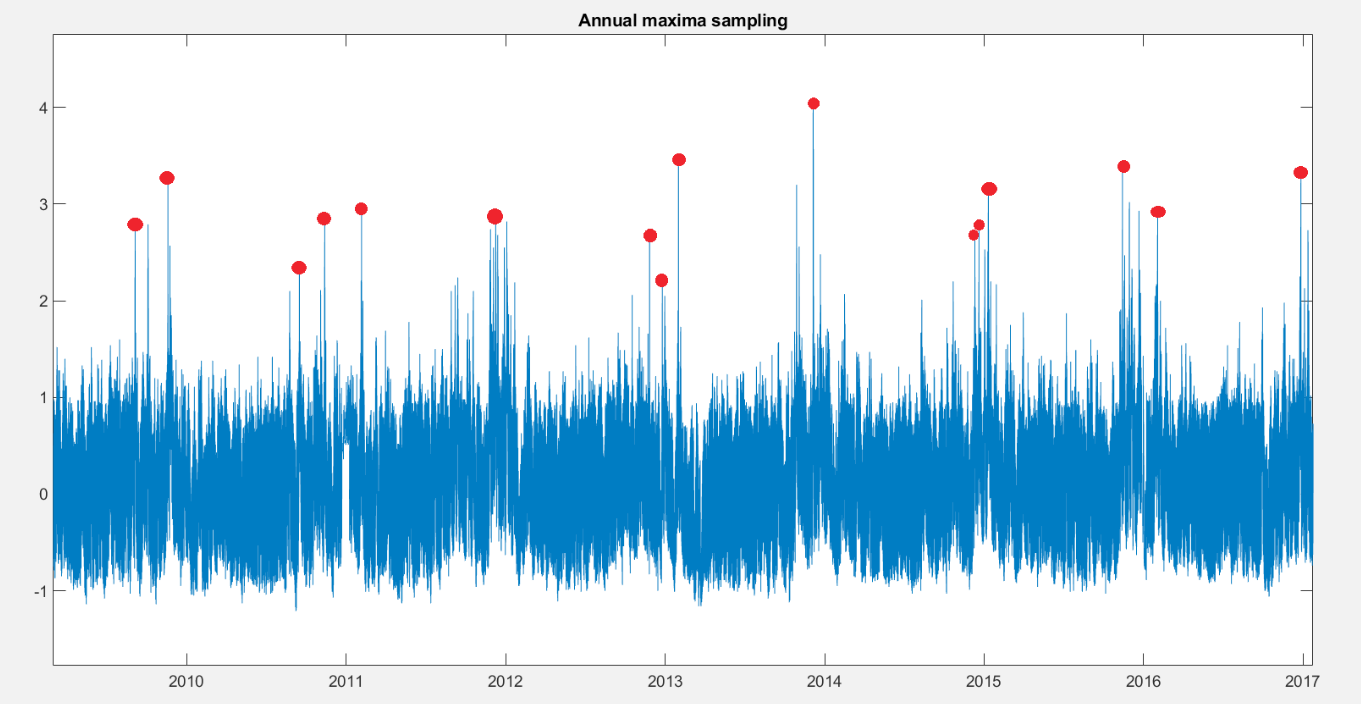


Figure 5 Illustration of the annual sampling. The two annual water level highs are chosen for the statistics

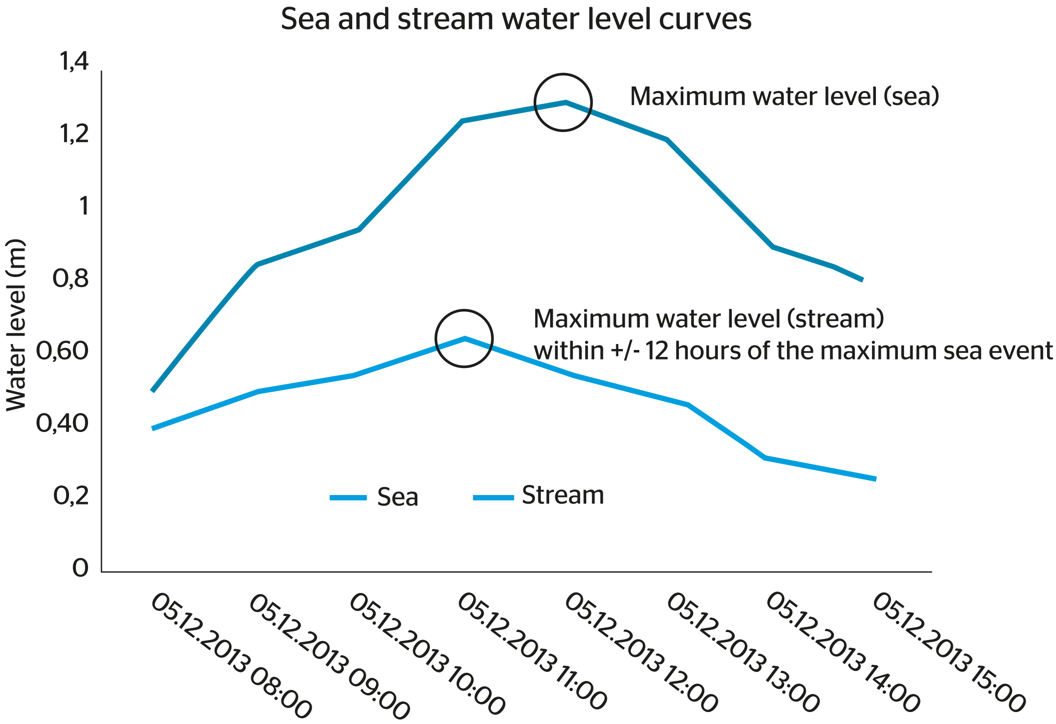


Figure 6 Illustration of the sampling method in regard to the +/- 12 hours requirement

4.4 Univariate Statistical Analyses

The univariate analysis is one of the first steps in the joint probability calculations. This is where the marginal distribution is found for each element.

When dealing with extreme data, extreme value distributions are used. The extreme value distribution, however, does not describe the behavior of a random variable but simply the tail of the distribution. Extreme statistics are usually tail statistics. For more information on this, please refer to (Gomes and Guillou, (2016)).

4.4.1 Empirical Distribution Functions

An Empirical distribution function (EDF) and the empirical cumulative distribution function (E-CDF) are both probability models for random data samples. Where the EDFs baseline is in the empirical data, the CDF is a hypothetical distribution model.

The EDF and CDF are both part of order statistics, which are the basis of extreme value statistics.

All elements within a sample are sorted going from smallest x and onward. By sorting the equidistant elements the chronological order of the data is lost. During this phase it is assumed that all elements from the sample are independent from one another.

When calculating the probability of exceedance, the likelihood that an event of a certain magnitude will occur becomes known. A general formula for various distribution functions and probability network (logarithmic, non-logarithmic, etc.) is provided below.

The general formula for EDF is:

$$\boldsymbol{C(u,v) = \frac{m-a}{n+1-2a}} \qquad \textbf{Equation 1}$$

Where:
C(u,v) is the non-exceedance probability,
a is the plotting position,
m is the rank, and
n is the sample size.

4.4.2 Annuality

Annuality, also known as the return interval (T), is the average time of which a certain value is exceeded. The empirical annuality can be determined from the empirical non-exceedance probability:

$$\boldsymbol{Exceedance\ probability = 1-(1-P)^{n}} \qquad \textbf{Equation 2}$$

Above the exceedance of probability was introduced, now a return period is attached to the size of a given event:

$$\boldsymbol{Return\ Interval\ (T) = \frac{1}{Exceedance\ probability}} \qquad \textbf{Equation 3}$$

Where:
P is the exceedance probability
n is the given year(s)

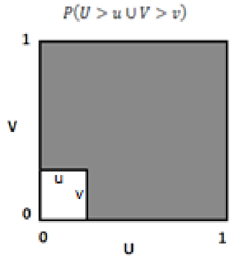
Estimates of extreme event return periods can be made with relatively short records. However, the associated confidence level in the flood frequency statistics is much higher with a longer period of data.

As an example, one would need 90 years on record to estimate a 10-year flood with no more than a ±10 % error. With a ±25 percent error, only 18 years on record is needed. It becomes clear that the length of data record needs to be within either ±10 % or ±25 % errors for the 20-, 50-, and 100-year floods. This is why it is not recommended to rely strongly on statistics made from data lengths shorter than twice the desired return period.

4.4.3 Return-plots

Determining which marginal distribution function to use for sea and stream depends strongly on the empirical data. The most efficient method to make this decision is to study the return-plot (see p. 13) of the two variables (sea and stream). The mathematical principles and the formula needed to find a return value for specific periods are stated below. The return intervals are separated into two categories: Less or Equal and Exceeding.

Less or Equal



$$C(u,v) = P(U \leq u, V \leq v)$$ Equation 4

This describes the probability that both realizations (U and V) are less than or equal to the value u and v. C(u,v) is the bivariate copula.

The corresponding probability of exceedance is calculated as follows:

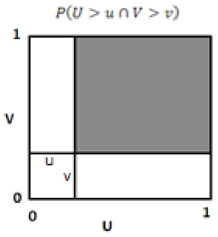
$$P(U > u \cap V > v) = 1 - u - v + C(u,v)$$ Equation 5

The return interval (T) is determined by:

$$T_{U \cap V} = \frac{\mu}{1-u-v+C(u,v)}$$ Equation 6

Where:
μ is the return period.

Exceeding



The probability that U or V will be exceeded is found as:

$$P(U > u \cup V > v) = 1 - C(u,v)$$ Equation 7

The return interval (T) is determined as follows:

$$T_{U \cup V} = \frac{\mu}{1-C(u,v)}$$ Equation 8

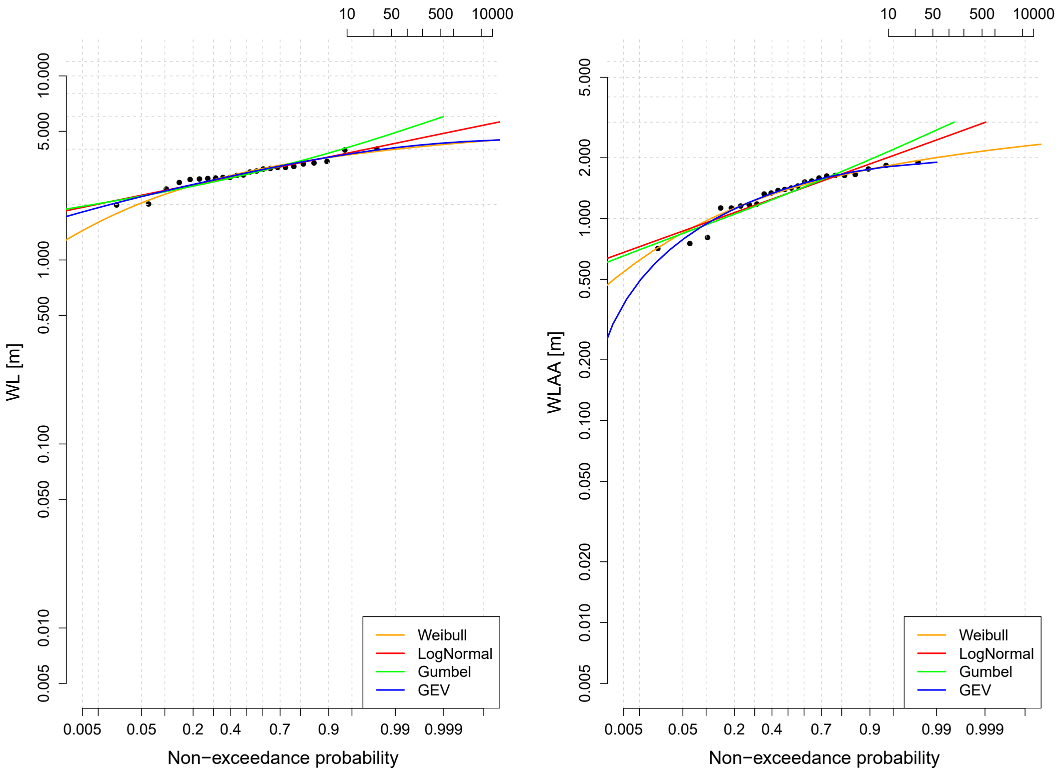


Figure 7 Return time plots for the sea (left) and a stream (right). Finding the marginal distribution is done by estimating parameters but also by examining the return time plot.

4.4.4 Extreme Value Distribution Functions

Extreme value theory is a separate branch of statistics that only deals with extreme events. The theory is based on the “extremal types” theorem, also referred to as the three types theorem.

This theorem states that only three types of distributions are required to model the maximum or minimum of the sample collection of random observations from the same distribution http://www.mathwave.com/articles/extreme-value-distributions.html#evd_gev. The distribution will show the trend and expected behavior of the individual data sources.

Presented below are the extreme value distributions that have been tested for the marginal distributions in the joint probability calculations in the presented work.

Weibull:

$$F_{(x)} = 1 - e^{-\left(\frac{x_i}{\beta}\right)^{\alpha}} \qquad \textbf{Equation 9}$$

Where:
α is the location parameter,
β is the shape parameter, and
η is the scale parameter.

LogNormal:

$$F_{(x)} = \frac{1}{2} ERFC\left(\frac{\mu - logx}{\sqrt{2}\sigma}\right) \qquad \textbf{Equation 10}$$

Where:
ERFC(x) returns the error function integrated between x and infinity,
μ is the mean of the associated normal distribution, and
σ is the standard deviation (of the associated normal distribution).

Gumbel:

$$\boldsymbol{F_{(x)} = e^{-e^{\frac{x_i-\alpha}{\beta}}}} \qquad \textbf{Equation 11}$$

Where:
α is the location parameter and
β is the shape parameter.

Other often-used extreme value distributions are the Generalized Extreme Value distribution (GEV) and the Pareto distribution. They are both very popular but proved less useful in this work due to the fact that the samples contain very few actual extremes. They are not used in the statistical analyses and are therefore not shown by formula.

4.4.5 Parameter Fitting – LMOM

For the parameter fitting the L- Moments (lmom) method is used. Lmoms are summary statistics for probability distributions and data samples that estimate the capability of the chosen distribution function to fit the empirical data sample. L-moms are analogous to the ordinary moments which measure the location, scale and shape of probability distributions or data samples. The difference between lmom and mom is that lmoms are computed from linear combinations of the ordered data values (see Vogel, R. M., and Wilson, I. (1996) for a more detailed description).

For this particular study lmoms are used because of less sensitivity to outliers in the data. (Because some data series are short, there are few or no actual extremes. However, when an extreme value is present, it will appear as an outlier, compared to the rest of the sample. This "outlier" is actually the only "actual" extreme value but because it is only one value, the rest of the sample values are used for the modeling).

A noticeable quality is that lmoms are more or less unbiased for all combinations of sample sizes and populations.

In brief: the lmom abilities are useful for providing accurate quantile estimates of the hydrological data where sample sizes are small (short time series) as in the case of stream data in this project.

The lmoms of the distribution function are calculated as follows:

$$\boldsymbol{\lambda_r = \int_0^1 x(F) P^*_{r-1}(F) df} \qquad \textbf{Equation 12}$$

Where:
x(F)is the inverse function of the CDF

$$\boldsymbol{P^*_r(F) = \sum_{k=0}^{r} (-1)^{r-k} \binom{r}{k} \binom{r+k}{k} F^k} \qquad \textbf{Equation 13}$$

Where:
Pr*(F) is the rth shifted Legendre polynomial

The lmoms of the sample are calculated as follows:

$$\boldsymbol{l_{r+1} = \sum_{k=0}^{r} p^*_{r,k^b k}} \qquad \textbf{Equation 14}$$

Where:
bk is the estimator of probability weighted moments

$$\boldsymbol{p^*_{r,k^b=(-1)^{r-k}\binom{r}{k}\binom{r+k}{k}}} \qquad \textbf{Equation 15}$$

$$\boldsymbol{b_k = n^{-1} \binom{n-1}{k}^{-1} \sum_{j=k+1}^{n} \binom{j-1}{k} x_{j:n}}$$ **Equation 16**

$$\boldsymbol{t_r = \frac{l_r}{l_2}, \quad r = 3, 4 \ldots}$$ **Equation 17**

4.4.6 Goodness of Fit - Goodness Testing

To test the correspondence between the theoretical distribution function and the empirical distribution estimated by plotting position, two tests are performed:

I The Graphical test (visual)

The correspondence between a distribution function and the sample values are checked visually. This is also used to identify outliers and it allows the viewer to assess the extrapolation behavior of the particular distribution function.

One objection to this method is that it is a subjective assessment, and it is therefore recommended to also perform a KS-test (see below).

II The Mathematical KS-test (Kolmogorov-Smirnov test)

The KS-test variable is equal to the maximum distance between the step function (obtained from the empirical distribution of the empirical values) and the theoretical distribution function. In this way a comparison between the theoretical and empirical distributions becomes available: the smaller the value, the better the fit.

4.5 Bivariate Statistical Analyses

4.5.1 Dependency

Certain dependencies between sea level and stream level are expected where the sea and stream meet. Dependence between sources is found by comparing the data series tendencies and the correlation between its coefficients.

Examples on types of dependence:

I Dependence (τ>0), is when the water level in the sea is rising, and then the water level in the stream rises as well

II Independent data sources are when the water level in the sea is rising but the stream data is non-influenced

Joint probability calculations for independent data example:

For independent data the probability of each variable (variable 1 and variable 2) can be multiplied to find the joint probability. An example is shown below:

1/25*1/50=0.0008
1/20 is the probability of the stream level of X value per year.
1/50 is the probability of the sea level of X value per year.

This means that there is a probability of 0.0008 (0.08%) that the sea and stream water levels combined will be equal to or higher than X.

III Negative dependence (τ<0) would be if the water level in the sea rose but the stream level dropped consistently.

When data sources are independent the joint probability of an occurrence is found by multiplying the marginal distributions with each other, however, with sources showing dependence the calculation of the joint probability becomes more complicated. To permit calculations of the probability of dependent sources, Copula functions are used.

Copulas are mathematical models that describe the dependency structures between two random variables independently from their marginal distributions. It links both univariate distributions via their dependency structure.

In order to measure the dependency two methods are used: Kendall's Tau and the Chi plot. For more information about the importance of dependence in Archimedean copulas refer e.g. to the Nelsen (1997) paper on Dependence and order in families of Archimedean copulas

4.5.1.1 Kendall's Tau and Kendall's Plot

This is a rank based measure of dependence. Each pair in the bivariate sample is split into concordant (consistent or in the same direction) and discordant (inconsistent – opposite direction) and counted:

$$\boldsymbol{\tau = \frac{Concordant - Discordant}{n(n-1)/2}} \qquad \textbf{Equation 18}$$

Where:
n is the number of pairs in the sample

Kendall's Tau is between -1 (negative dependence) and 1 (positive dependence). In the case of negative dependence (an example could be that the sea water level is increasing while the stream water level is decreasing), only the Frank Copula is a possible choice for calculating the joint probability. Positive dependence here is when the sea level rises and the stream level does too, as a consequence of the sea level rising.

An example of independent data and dependent data is shown in figure 8 below:

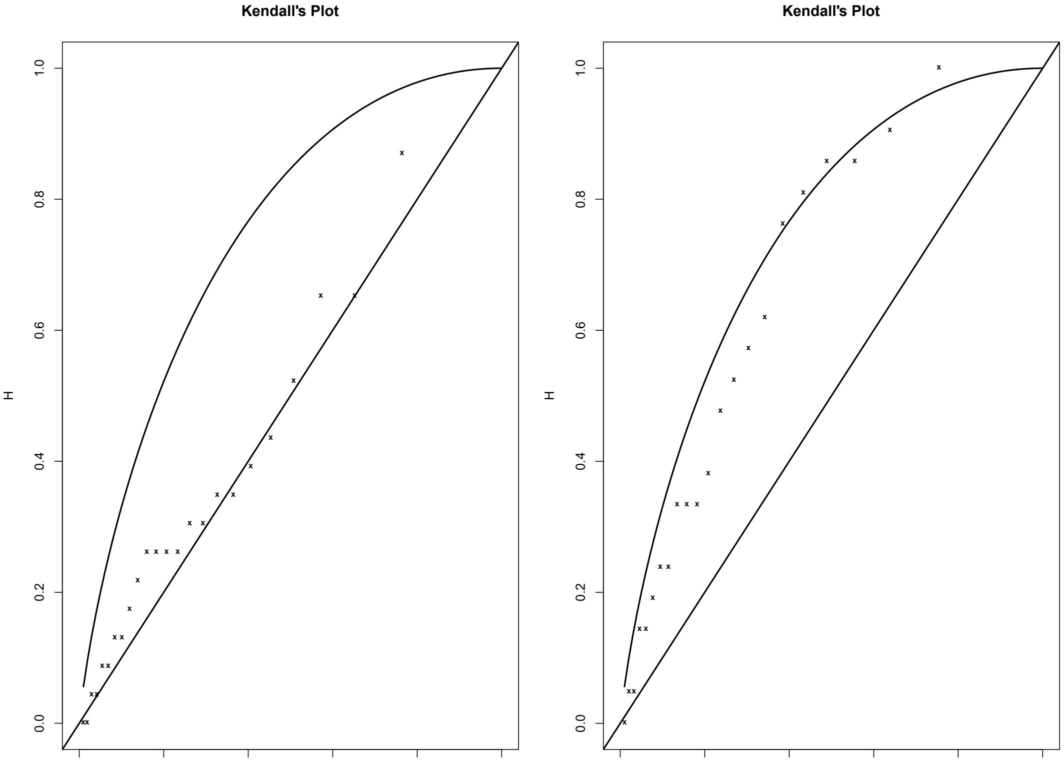


Figure 8 Independent data pairs are placed near or on the diagonal line Wi:n = Hi (left), whereas dependent data tend to follow the curved line (right)

4.5.1.2 Chi plot

The Chi plot is a graphical representation of the measures of local dependency. It contains more information regarding the usual measures of correlation.

The value λi on the x-axis is a measure of the distance of event pair (xi,yi) from the median of the two-dimensional sample. χi is a correlation coefficient between xi and yi.

When there is independence between the variables in the pair, χi is randomly distributed around zero. If variable 1 constantly follows variable 2 in increase, then χi =1. If variable 1 constantly follows variable 2 in decrease, then χi =-1. For further reading on the Chi-plot please see Marchi, V., Rojas, F. and Louzada, F., 2012.

It is generally recommended to fully appreciate the importance of dependence before any interpretation of data and results is attempted.

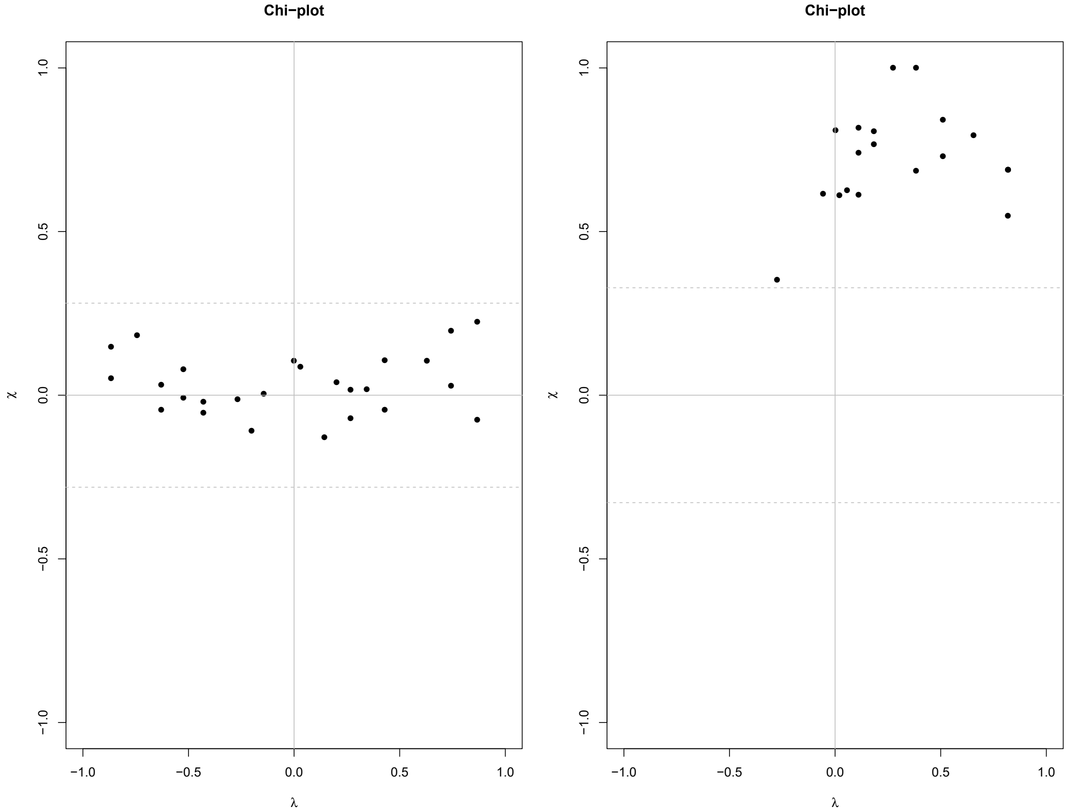


Figure 9 Independent data pairs are placed near zero (left), whereas mostly dependent data tend to go towards the value if there is a positive dependence 1 and -1 if there is a negative dependence (right)

4.5.1.3 The Archimedean Copula Functions

The relationship between a copula function C and the bivariate function:

$$F_{X,Y}(x,y) = C[F_X(x), F_Y(y)]$$ **Equation 19**

X and Y are both correlated random variables, which can be represented individually by their univariate marginal distributions FX(x) (also denoted as u) and FY(y) (also denoted as v).

u and v both represent a specific value of the correlated variable and range between O and 1.

The copula only contains information on the nature and extent of the dependency of the variables. There is no connection between the description of the dependence and the marginal distributions, and the marginal distributions do not contain any information on the correlation.

Another particular thing about copulas is that they can be understood as a bivariate distribution function analogous to a univariate distribution function; however their densities can assume values greater than 1, which can appear to be an error. It is not.

In the copula function it is the parameter theta (ϴ) that influences the degree of dependency. Equation 20-23 shows the most commonly used Archimedean copulas (3-parameters):

$$C(u,v) = e^{[-((-log u)^{\theta}+(-log v)^{\theta})^{1/\theta}]} \qquad \theta \in [1,\infty) \qquad \text{Equation 20}$$

$$C(u,v) = -\frac{1}{\theta}\log[1+\frac{(e^{-\theta u}-1)(e^{-\theta v})}{e^{-\theta}-1}] \qquad \theta \in \mathbb{R}\{0\} \qquad \text{Equation 21}$$

$$C(u,v) = [\max\{u^{-\theta}+v^{-\theta}-1;0\}]^{-1/\theta} \qquad \theta \in [1,\infty)\backslash\{0\} \quad \text{Equation 22}$$

$$C(u,v) = 1-[(1-u)^{\theta}+(1-v)^{\theta}-(1-u)^{\theta}(1-v)^{\theta}]^{1/\theta} \qquad \theta \in [1,\infty) \qquad \text{Equation 23}$$

Some of the two parameter Archimedean copulas are:

BB1 = Clayton-Gumbel, BB6 = Joe-Gumbel, BB7 = Joe-Clayton and BB8 = Joe-Frank.

4.5.1.4 Parameter fitting

The parameters of the copulas are estimated using the lmom. See section 4.4 for a full description.

4.5.1.5 Goodness of fit test – Graphical test

When it comes to choosing the right copula function to represent data, the first part of the process is to locate the better match between the observed (black) and the generated (blue) graph, see figure 10 below. The different copula functions (Clayton, Dumbel-Hougaard, Frank, etc.) are plotted onto the observed data. A good fit in this case would be the Clayton or the BB7 function. A poor fit would be Frank.

When two copula functions are visually close in fit, it is advised to consider the tau and theta values for determining the overall best fit: the lower the values the better the fit. These are the coefficients describing the quality, or, goodness of the fit.

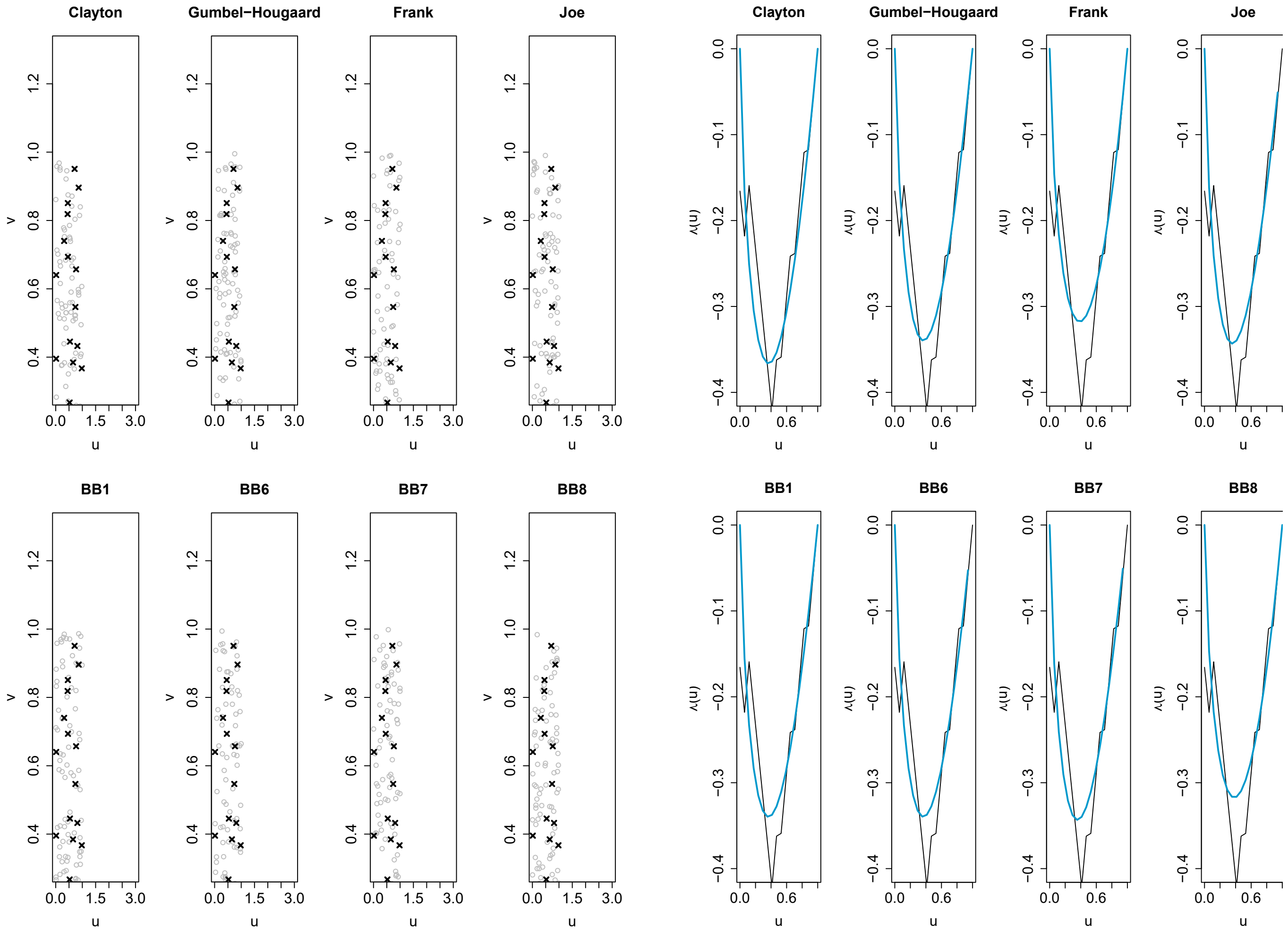


Figure 10 Left - scatterplots of the individual copula functions. The theoretical data dots (gray circles) should imitate the empirical data (black dots). If this is the case then the model shows potential as the chosen parameter, however there are several graphical tests thatshould be taken into account. To the right - the copula functions (blue graphs) are individually fit in comparison to the transformed empirical data (black graphs).

4.5.1.6 The Copula plot

The copula plot creates a visualization of the combined events and the combination probability.

The X-axis shows the detrended sea water level and the Y-axis is the water level and time corresponding stream data. The plotted data are the two annual maxima sea values from all years in the data series. There is the possibility of some stations having fewer data points, in these cases it is due to outlier removal. Because only sea data dictates the level, some stream water level may be very low or non-existing (error measurement).

The three curves are isolines of the return time 25-, 50- and 100-years (MT25, MT50 and MT100) which gives the probability of occurrence of the combined events.

On the plot, 3 circles are drawn (1-orange, 2-green and 3-blue) on the MT100 isoline. These circles represent three scenarios of combinations resulting in a combined MT100 water level.

- Circle 1 is a combination of high stream level and low sea level.
- Circle 2 is a combination of medium sea and stream water level.
- Circle 3 is a combination of high sea level and low stream level.

Depending on the area, high sea water level may play a bigger role, in regard to flooding, in comparison to high stream water level or the other way around.

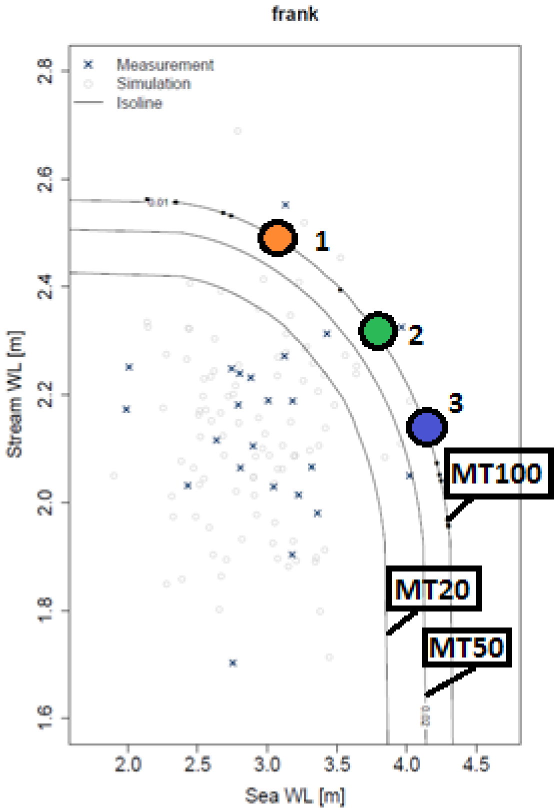


Figure 11 shows an example of a Copula plot explained.

5. Calculating the Probability of Sea and Stream Maximum Occurring on the same Day

To calculate the probability of maximum water level both in sea and stream on the same day, the number of days in the hydrological season (storm season) need to be determined.Because annual maximum events are known to occur in September, this month is included in the calculations.

1 There are 212 days (30+31+30+31+31+28+31) in the hydrological season (from the 1st of September to the 31st of March).

2 To calculate the probability there are these assumptions:

a. Theduration of the storm event isonly one day (sea)

b. The duration of the flood event is only one day (streams)

c. All the events are equally distributed (Uniform joint distribution)

3 The probability that a maximum event occurs on a day during the storm season is: $\frac{1}{212} = 0.0047$. This is valid for sea and streams alike.

4 The probability that a maximum event occurs in both sea and stream on day 1 of the storm season is: $\frac{1}{212} * \frac{1}{212} = 0.000022\ (or\ 0.0022\%)$

5 The probability that a maximum event occurs in both sea and stream on day 1 or day 2 or ... day 212 using the "AM" method with – 1 event/year is:

$$\underbrace{\frac{1}{212} * \frac{1}{212}}_{\text{(day 1)}} + \underbrace{\frac{1}{212} * \frac{1}{212}}_{\text{(day 2)}} +.. \underbrace{\frac{1}{212} * \frac{1}{212}}_{\text{(day 212)}} = \Sigma_1^{212} \frac{1}{212} * \frac{1}{212} = 212\left(\frac{1}{212} * \frac{1}{212}\right) = 0.0047\ (0.47\%)$$

6 Calculating a scenario of two events per year is slightly more complicated. One has to imagine a scenario of: (Max sea #1 and max stream #1) or (max sea #2 and max stream #2) or (max sea #1 and max stream #2) or (max sea #2 and max stream #1)

The probability that a maximum event occurs in both sea and stream on day 1 or day 2 or ... day 212 using the "AM" method with – 2 events/year is:

$$\sum_{1}^{212} \frac{1}{212} * \frac{1}{212} \quad + \quad \sum_{1}^{211} \frac{1}{211} * \frac{1}{211} \quad + \quad \sum_{1}^{212} \frac{1}{212} * \frac{1}{212} \quad + \quad \sum_{1}^{211} \frac{1}{211} * \frac{1}{211} \quad = 0.019\ (or\ 1.9\%)$$

(Max sea #1 and max stream #1) or (max sea #2 and max stream #2) or (max sea #1 and max stream #2) or (max sea #2 and max stream #1)

! Notice how one day is "lost" in the second event!

6. Results – Joint Probability

The joint probability of extreme water levels in sea and stream is calculated and presented for all stations below.

Station	Station name (in R)	Method - Block maxima	Distribution function SEA	Distribution function STREAM	Copula function	Fit (τ og Θ)
Ribe Å, Kammerslusen	Ribe 1	”Re” 2/yr	LogNormal	LogNormal	Clayton	0.2640693 - 0.4257247
Ribe Vester Å, Ribe Havn	Ribe 2	”Re” 2/yr	LogNormal	LogNormal	Clayton	0.2173913 - 0.3893619
Ribe Å	Ribe 3	”Re” 2/yr	LogNormal	LogNormal	Frank	-0.01185771 - (-0.08879279)
Ribe Å, Stavnager	Ribe 4	”Re” 2/yr	LogNormal	LogNormal	Clayton	0.1304348 - 0.1048983
Hjortvad Å	Ribe 5	”Re” 2/yr	LogNormal	LogNormal	BB1	0.1503268 - 0.0010000

Table 2 Overview of stations ID, respective marginal distributions, copula functions and the copula fit.

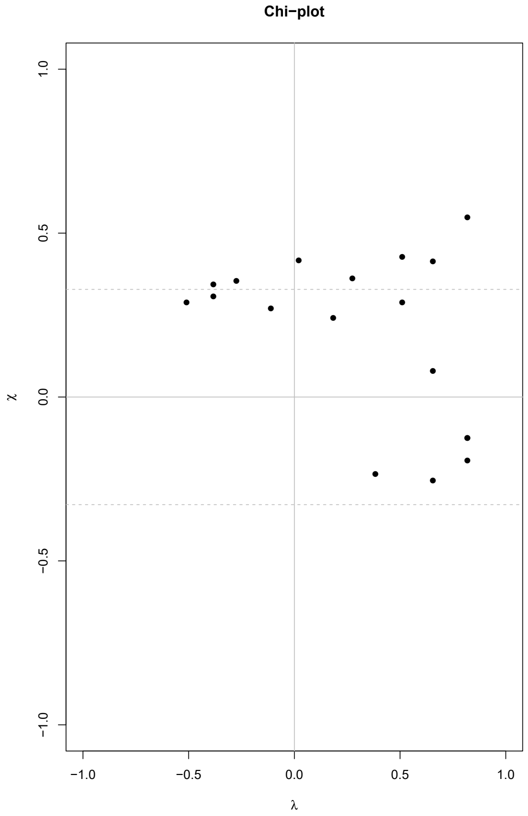
6.1 Ribe 1

6.1.1 Sample

1. value pr. year Date	Hydro. year	Sea (m)	Stream (m)	2. value pr. year Date	Hydro year	Sea (m)	Stream (m)
08-01-2005 16:00	2005	3.96	1.52	18-11-2004 05:00	2005	3.18	1.05
26-10-2005 07:00	2006	2.01	1.01	15-11-2005 01:00	2006	1.99	0.73
01-03-2008 19:00	2008	3.32	1.53	01-02-2008 09:00	2008	3.05	1.61
04-09-2009 13:00	2009	2.76	0.63	10-11-2008 11:00	2009	2.43	1.15
18-11-2009 15:00	2010	3.22	1.07	04-10-2009 03:00	2010	2.75	0.70
05-02-2011 04:00	2011	2.90	1.06	12-11-2010 19:00	2011	2.81	1.25
09-12-2011 12:00	2012	2.80	1.64	03-01-2012 22:00	2012	2.79	1.56
31-01-2013 03:00	2013	3.43	1.38	25-11-2012 23:00	2013	2.64	1.07
05-12-2013 16:00	2014	4.02	1.35	28-10-2013 16:00	2014	3.18	1.27
11-01-2015 04:00	2015	3.13	1.58	20-12-2014 13:00	2015	2.75	1.47
14-11-2015 04:00	2016	3.36	1.28	30-11-2015 03:00	2016	3.01	1.52

Table 3 shows the sample from Ribe 1

6.1.2 Chi- plot and Kendall's plot for dependence



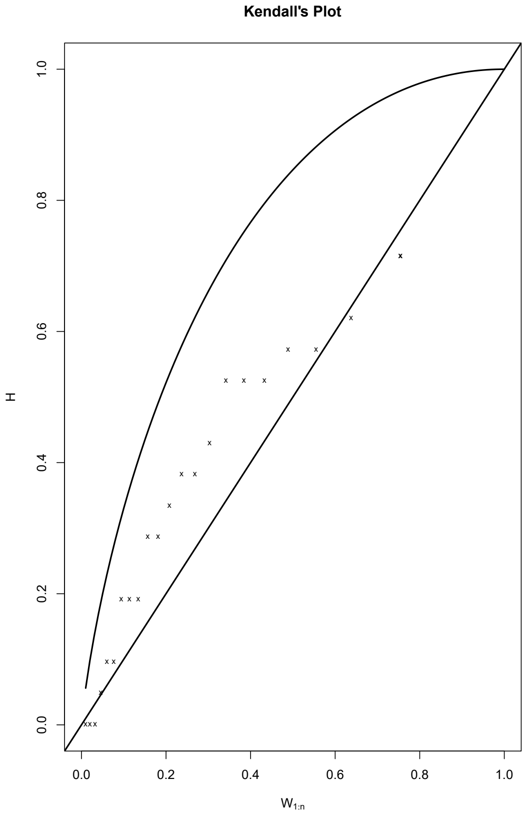
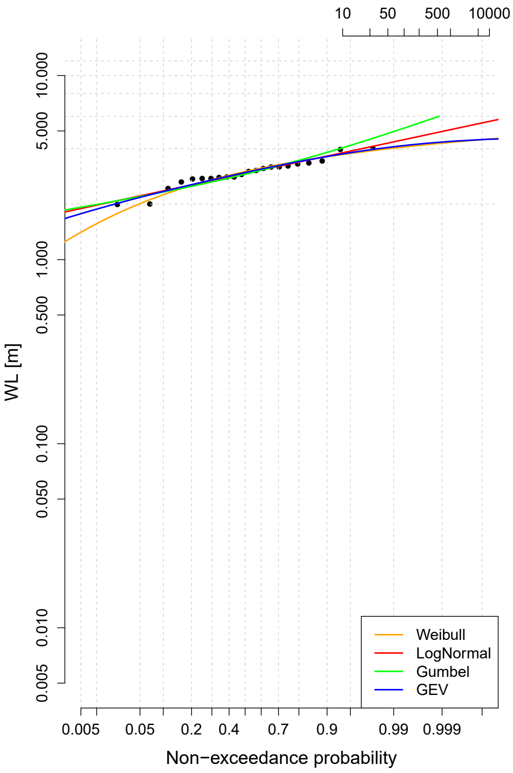


Figure 12 Ribe 1's Chi-plot

Figure 13 Ribe 1's Kendall's plot)

6.1.3 Return time plots for Sea (left) and Stream (right)



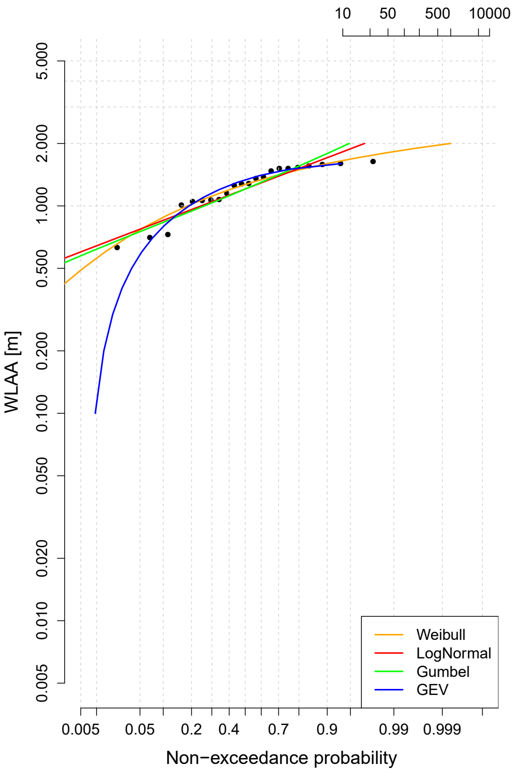


Figure 14 Return plot for Ribe (1) sea

Figure 15 Return plot for Ribe 1 (stream)

6.1.4 Copula fitting: Scatter- and lambda -plot

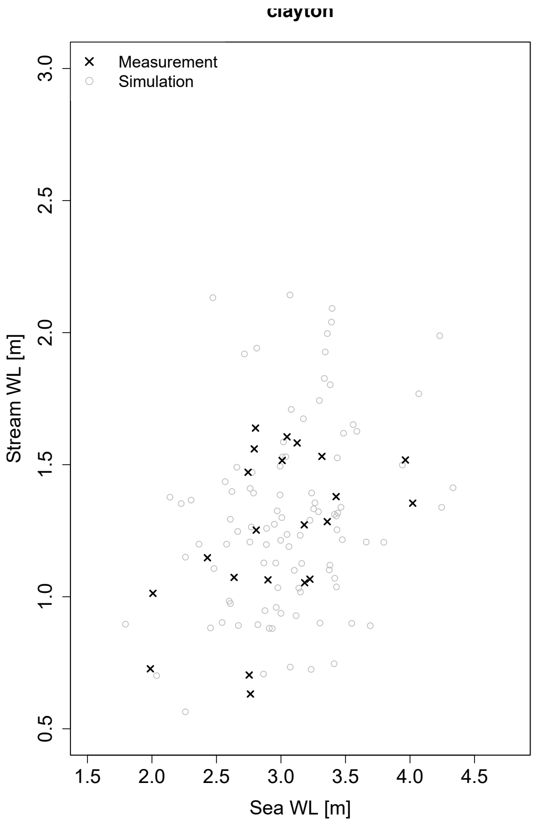


Figure 16 Scatter plot for Ribe 1 data

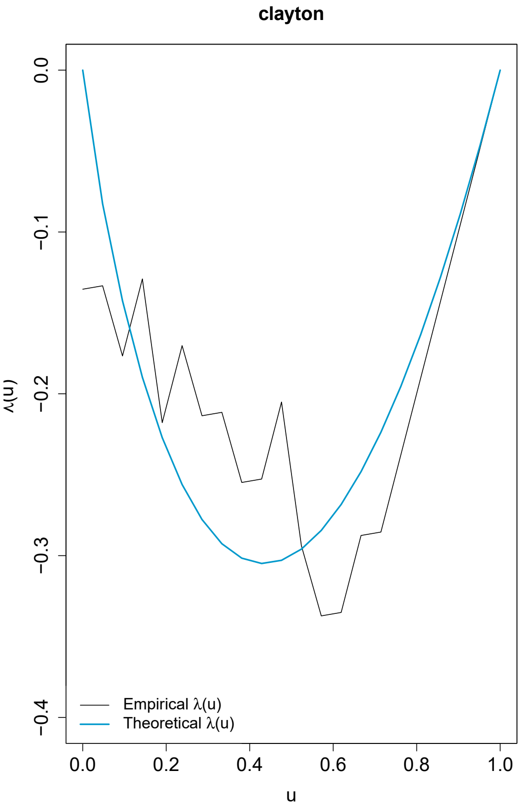


Figure 17 Lambda copula plot for Ribe 1

6.1.5 Copula plot

With 20 random pairs plotted on the 100 year return interval isoline (left). The 20 pair sample is shown in the table to the right.

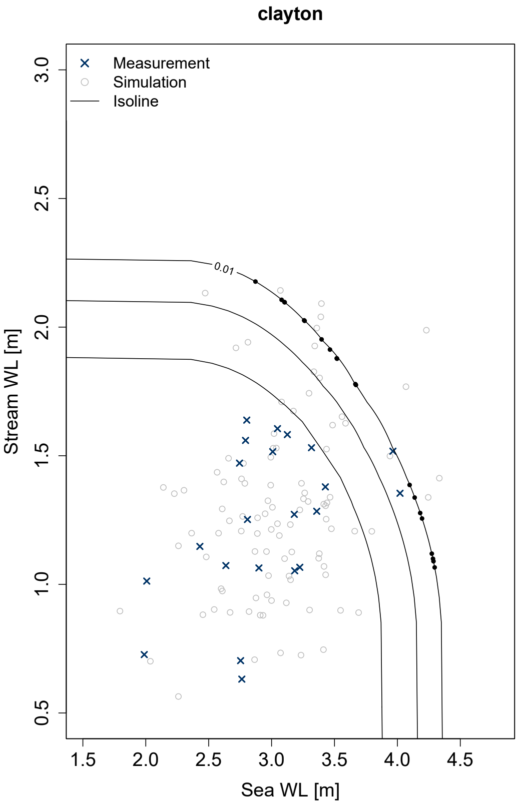


Figure 18 shows the copula plot from Ribe 1

Pair	Sea (m)	Stream (m)	Pair	Sea (m)	Stream (m)
1	4.27	1.12	11	4.18	1.28
2	2.87	2.18	12	4.10	1.39
3	3.52	1.88	13	3.26	2.03
4	3.10	2.10	14	4.14	1.34
5	3.52	1.88	15	4.29	1.09
6	3.67	1.78	16	3.08	2.11
7	4.28	1.10	17	4.30	1.07
8	3.46	1.91	18	3.40	1.95
9	3.26	2.02	19	4.20	1.26
10	3.67	1.78	20	3.10	2.10

Table 4 shows the 20 random pairs from isoline MT100

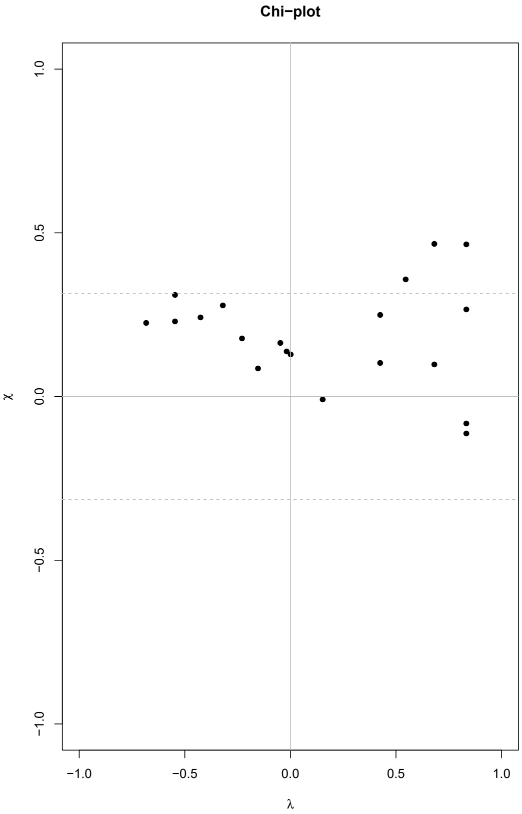
6.2 Ribe 2

6.2.1 Sample

1. value pr. year	Hydro.	Sea (m)	Stream	2. value pr. year	Hydro	Sea (m)	Stream
Date	year		(m)	Date	year		(m)
08-01-2005 16:00	2005	3.96	1.66	18-11-2004 05:00	2005	3.18	1.13
26-10-2005 07:00	2006	2.01	1.13	15-11-2005 01:00	2006	1.99	0.81
12-01-2007 07:00	2007	3.13	1.90	18-03-2007 15:00	2007	2.88	1.42
01-03-2008 19:00	2008	3.32	1.52	01-02-2008 09:00	2008	3.05	1.59
04-09-2009 13:00	2009	2.76	0.71	10-11-2008 11:00	2009	2.43	1.18
18-11-2009 15:00	2010	3.22	1.15	04-10-2009 03:00	2010	2.75	0.75
05-02-2011 04:00	2011	2.90	1.39	12-11-2010 19:00	2011	2.81	1.45
09-12-2011 12:00	2012	2.80	1.76	03-01-2012 22:00	2012	2.79	1.62
31-01-2013 03:00	2013	3.43	1.53	25-11-2012 23:00	2013	2.64	1.17
05-12-2013 16:00	2014	4.02	1.38	28-10-2013 16:00	2014	3.18	1.33
11-01-2015 04:00	2015	3.13	1.83	20-12-2014 13:00	2015	2.75	1.64
14-11-2015 04:00	2016	3.36	1.34	30-11-2015 03:00	2016	3.01	1.64

Table 5 shows the sample from Ribe 2

6.2.2 Chi- plot and Kendall's plot for dependence



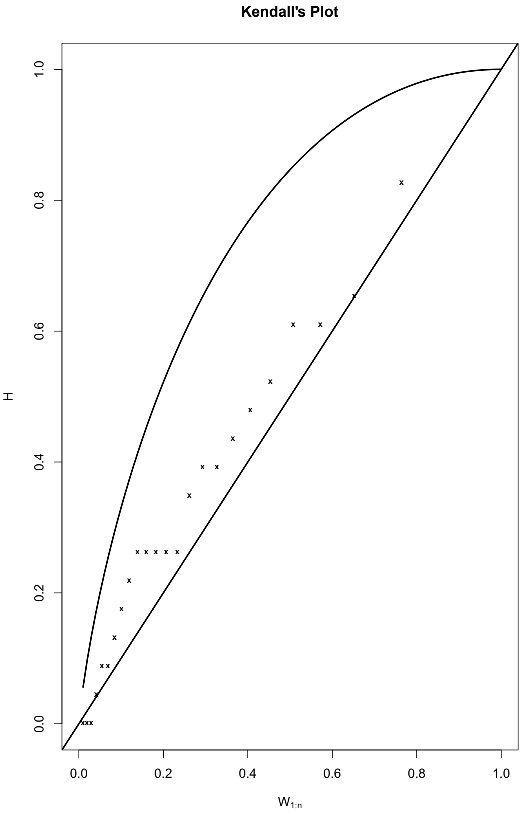


Figure 19 Ribe 2's Chi-plot

Figure 20 Ribe 2's Kendall's plot1

6.2.3 Return time plots for Sea (left) and Stream (right)

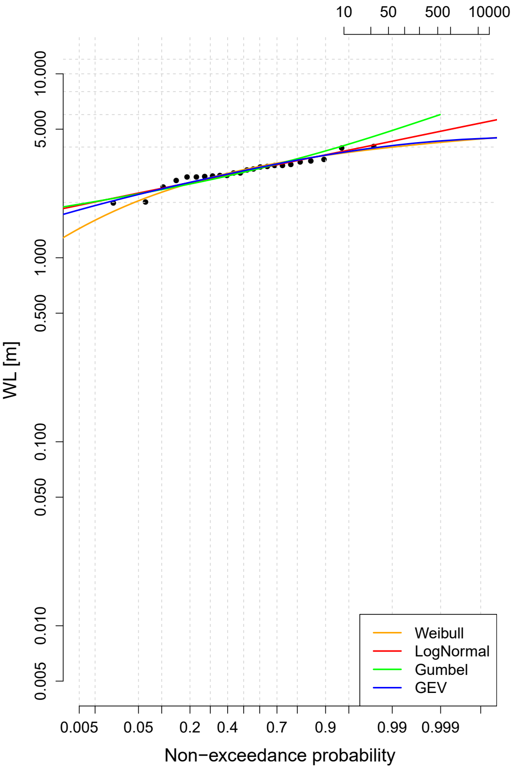


Figure 21 Return plot for Ribe sea (2)

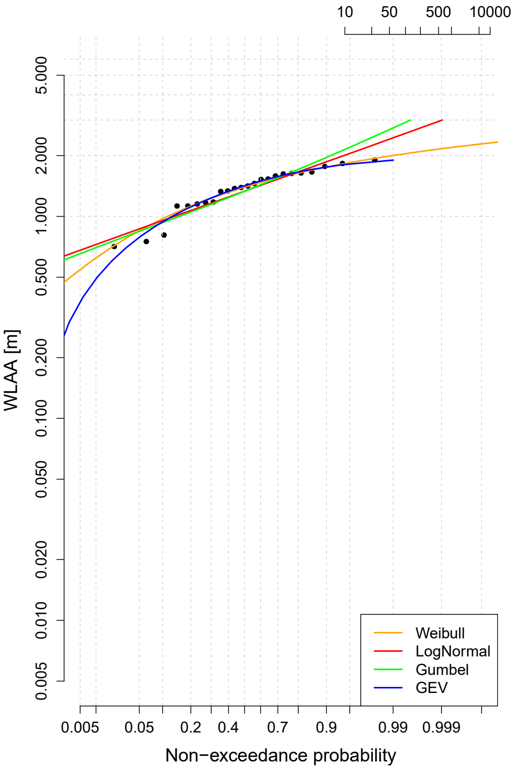
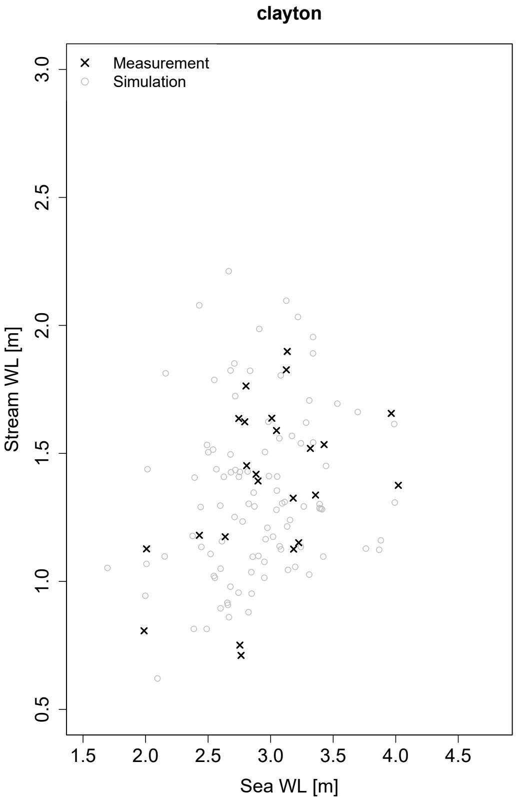


Figure 22 Return plot for Ribe 2 (stream)

6.2.4 Copula fitting: Scatter- and lambda -plot



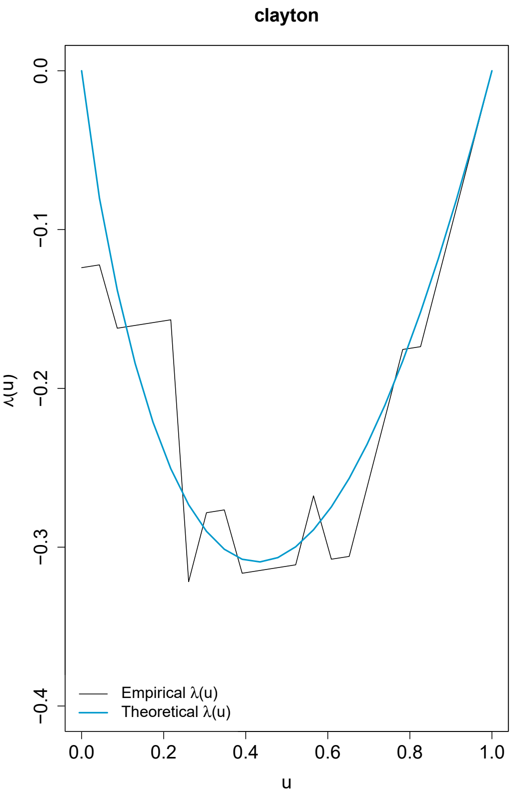


Figure 23 Scatter plot for Ribe 2 data

Figure 24 Lambda copula plot for Ribe 2

6.2.5 Copula plot

With 20 random pairs plotted on the 100 year return interval isoline (left). The 20 pair sample is shown in the table to the right.

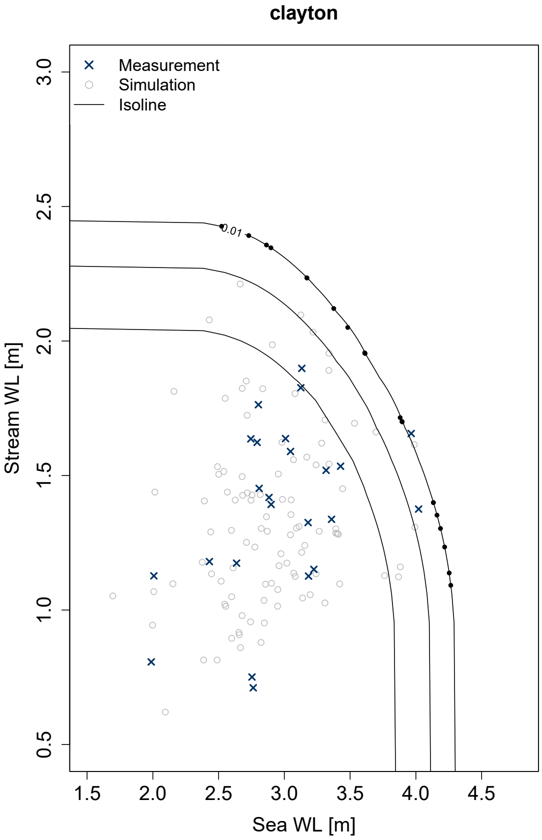


Figure 25 shows the copula plot from Ribe 2

Pair	Sea (m)	Stream (m)	Pair	Sea (m)	Stream (m)
1	4.13	1.40	11	2.86	2.36
2	4.13	1.40	12	3.48	2.05
3	3.17	2.23	13	3.17	2.23
4	4.25	1.14	14	3.89	1.70
5	2.90	2.35	15	3.61	1.95
6	3.90	1.70	16	2.73	2.39
7	4.16	1.35	17	2.52	2.43
8	3.38	2.12	18	4.19	1.30
9	4.22	1.23	19	4.27	1.09
10	3.61	1.96	20	3.88	1.72

Table 6 shows the 20 random pairs from isoline MT100 in the Ribe 2 copula plot

6.3 Ribe 3

6.3.1 Sample

1. value pr. year Date	Hydro. year	Sea (m)	Stream (m)	2. value pr. year Date	Hydro year	Sea (m)	Stream (m)
08-01-2005 16:00	2005	3.96	2.33	18-11-2004 05:00	2005	3.18	2.19
26-10-2005 07:00	2006	2.01	2.25	15-11-2005 01:00	2006	1.99	2.17
12-01-2007 07:00	2007	3.13	2.55	18-03-2007 15:00	2007	2.88	2.23
01-03-2008 19:00	2008	3.32	2.07	01-02-2008 09:00	2008	3.05	2.03
18-11-2009 15:00	2010	3.22	2.01	10-11-2008 11:00	2009	2.43	2.03
05-02-2011 04:00	2011	2.90	2.11	04-10-2009 03:00	2010	2.75	1.70
09-12-2011 12:00	2012	2.80	2.24	12-11-2010 19:00	2011	2.81	2.06
31-01-2013 03:00	2013	3.43	2.31	03-01-2012 22:00	2012	2.79	2.18
05-12-2013 16:00	2014	4.02	2.05	25-11-2012 23:00	2013	2.64	2.12
11-01-2015 04:00	2015	3.13	2.27	28-10-2013 16:00	2014	3.18	1.90
14-11-2015 04:00	2016	3.36	1.98	20-12-2014 13:00	2015	2.75	2.25
				30-11-2015 03:00	2016	3.01	2.19

Table 7 shows the sample from Ribe 3

6.3.2 Chi- plot and Kendall's plot for dependence

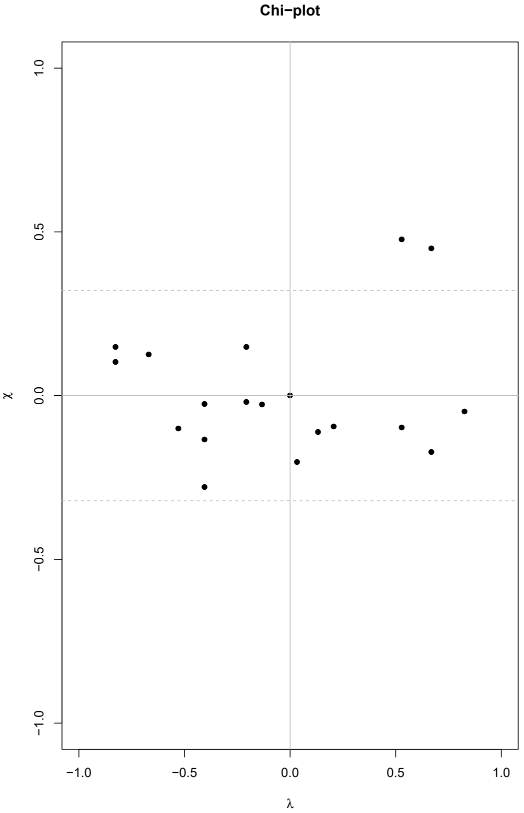


Figure 26 Ribe 3's Chi-plot

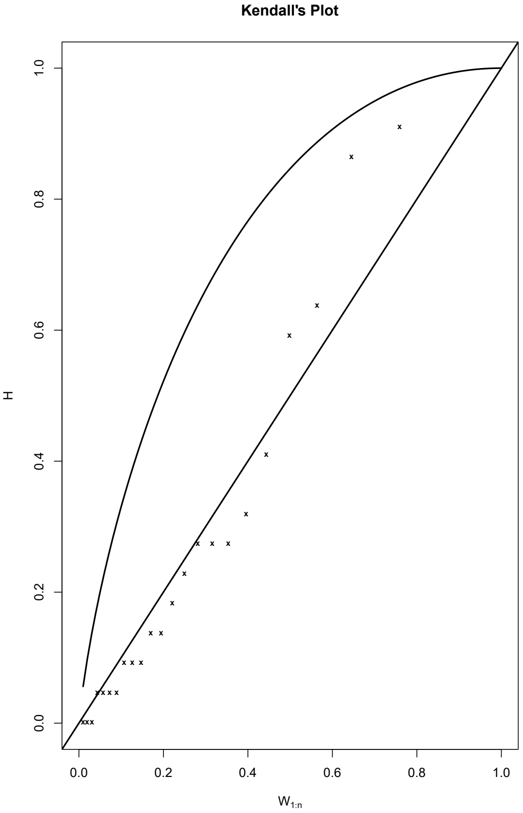


Figure 27 Ribe 3's Kendall's plot

6.3.3 Return time plots for Sea (left) and Stream (right)

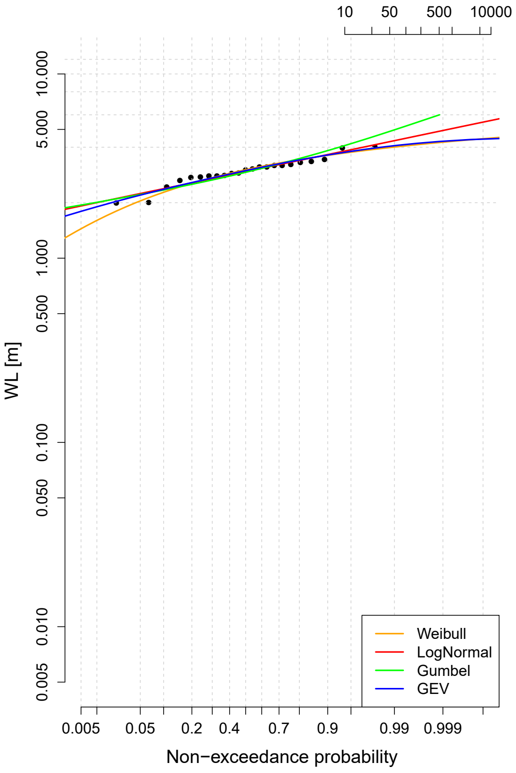


Figure 28 Return plot for Ribe sea (3)

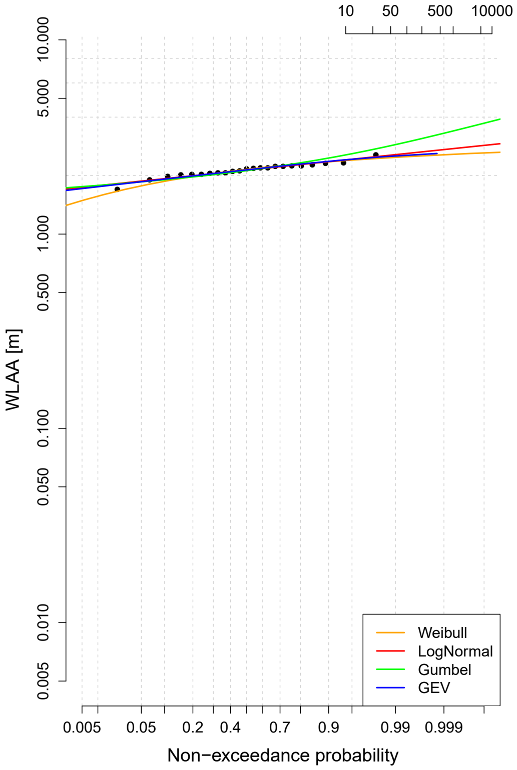
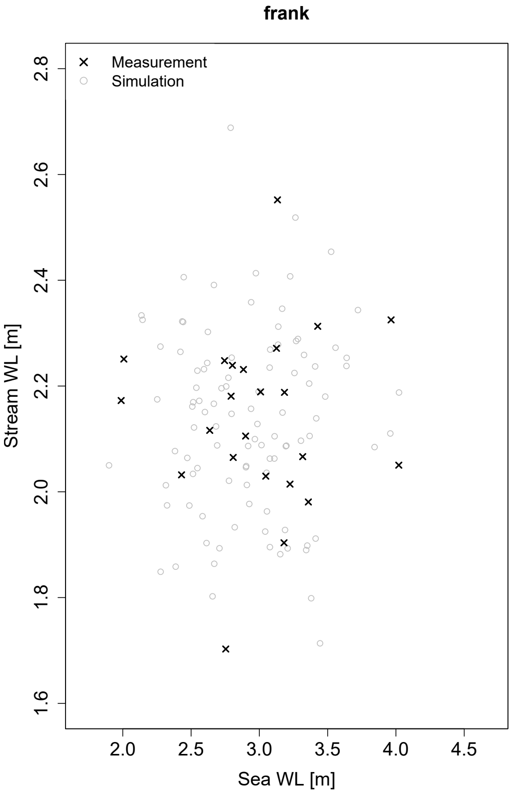


Figure 29 Return plot for Ribe 3 (stream)

6.3.4 Copula fitting: Scatter- and lambda -plot



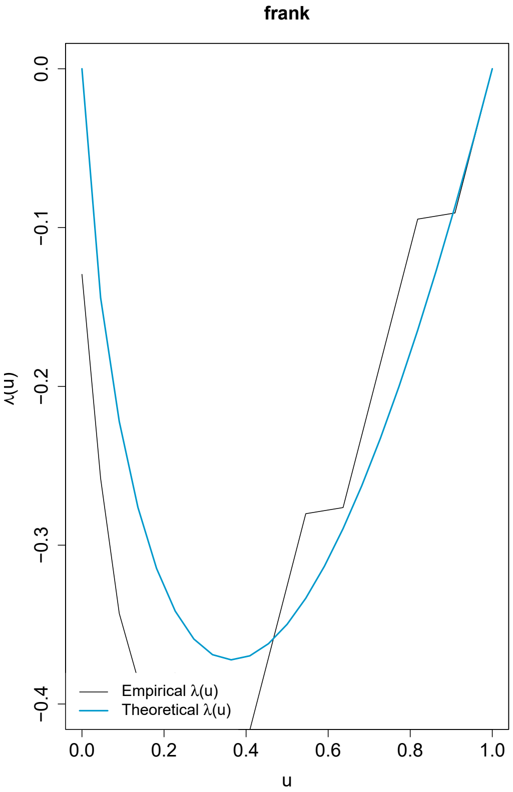


Figure 30 Scatter plot for Ribe 3 data

Figure 31 Lambda copula plot for Ribe 3

6.3.5 Copula plot

With 20 random pairs plotted on the 100 year return interval isoline (left). The 20 pair sample is shown in the table to the right.

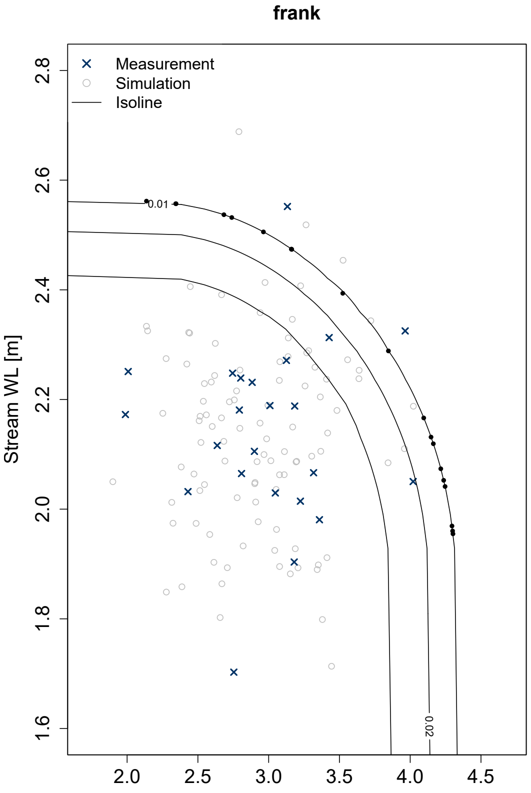


Figure 32 shows the copula plot from Ribe 3

Pair	Sea (m)	Stream (m)	Pair	Sea (m)	Stream (m)
1	4.30	1.97	11	2.96	2.51
2	4.25	2.04	12	4.30	1.96
3	3.16	2.47	13	3.52	2.39
4	4.10	2.17	14	2.74	2.53
5	4.24	2.05	15	4.16	2.12
6	3.16	2.47	16	4.15	2.13
7	2.34	2.56	17	2.68	2.54
8	4.30	1.95	18	3.85	2.29
9	2.14	2.56	19	4.22	2.07
10	2.35	2.56	20	3.16	2.47

Table 8 shows the 20 random pairs from isoline MT100 in the Ribe 3 copula plot

6.4 Ribe 4

6.4.1 Sample

1. value pr. year Date	Hydro. year	Sea (m)	Stream (m)	2. value pr. year Date	Hydro year	Sea (m)	Stream (m)
08-01-2005 16:00	2005	3.96	1.37	18-11-2004 05:00	2005	3.18	0.99
26-10-2005 07:00	2006	2.01	0.91	15-11-2005 01:00	2006	1.99	0.84
12-01-2007 07:00	2007	3.13	2.00	18-03-2007 15:00	2007	2.88	1.26
01-03-2008 19:00	2008	3.32	1.26	01-02-2008 09:00	2008	3.05	1.25
04-09-2009 13:00	2009	2.76	0.38	10-11-2008 11:00	2009	2.43	0.92
18-11-2009 15:00	2010	3.22	1.08	04-10-2009 03:00	2010	2.75	0.35
05-02-2011 04:00	2011	2.90	1.20	12-11-2010 19:00	2011	2.81	1.23
09-12-2011 12:00	2012	2.80	1.55	03-01-2012 22:00	2012	2.79	1.49
31-01-2013 03:00	2013	3.43	1.78	25-11-2012 23:00	2013	2.64	1.20
05-12-2013 16:00	2014	4.02	0.89	28-10-2013 16:00	2014	3.18	0.84
11-01-2015 04:00	2015	3.13	1.43	20-12-2014 13:00	2015	2.75	1.51
14-11-2015 04:00	2016	3.36	0.82	30-11-2015 03:00	2016	3.01	1.46

Table 9 shows the sample from Ribe 4

6.4.2 Chi- plot and Kendall's plot for dependence

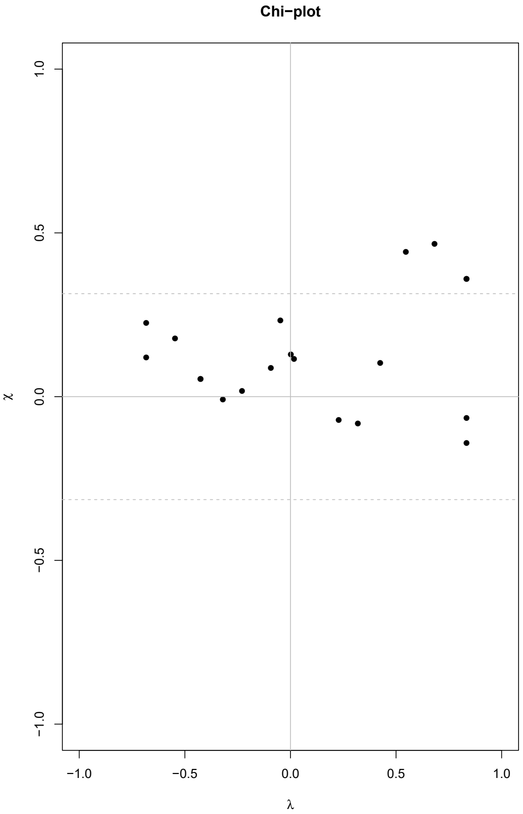


Figure 33 Ribe 4's Chi-plot

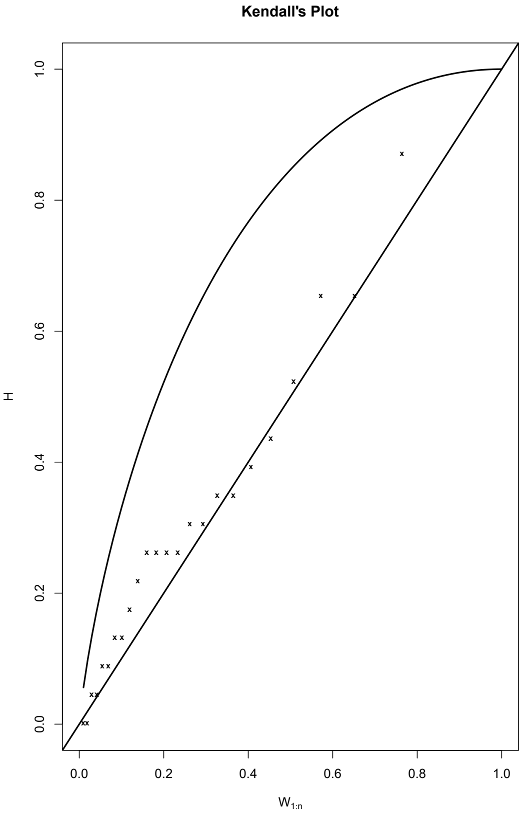


Figure 34 Ribe 4's Kendall's plot

6.4.3 Return time plots for Sea (left) and Stream (right)

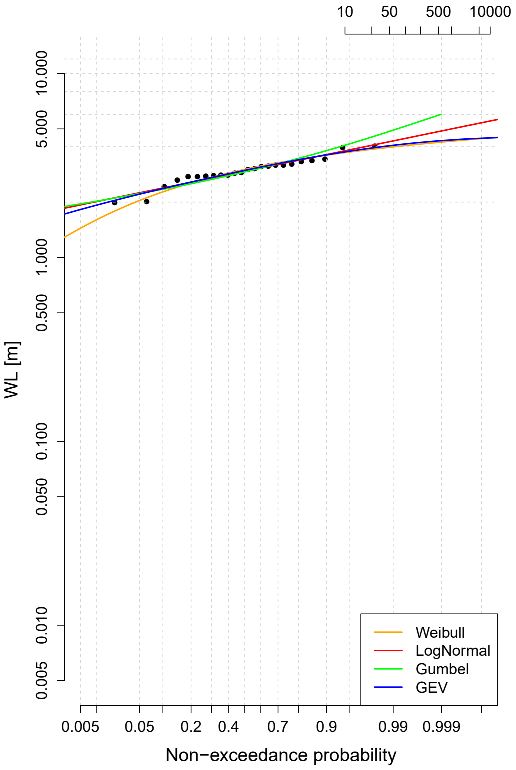


Figure 35 Return plot for Ribe sea (4)

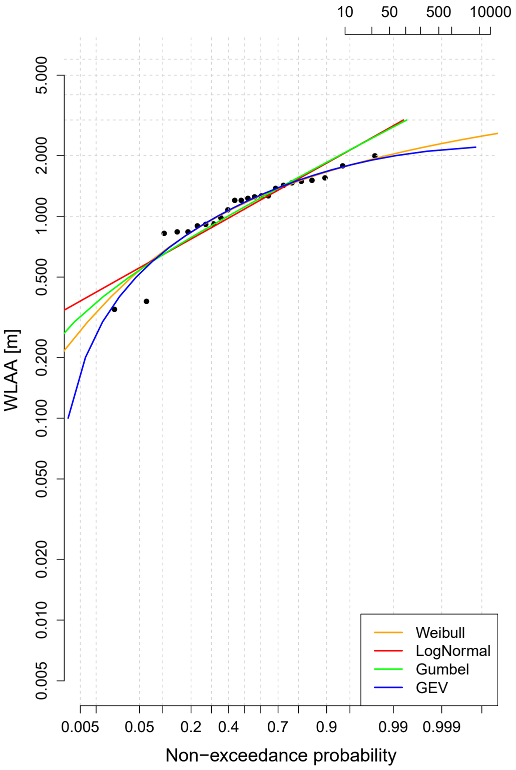


Figure 36 Return plot for Ribe 4 (stream)

6.4.4 Copula fitting: Scatter- and lambda -plot

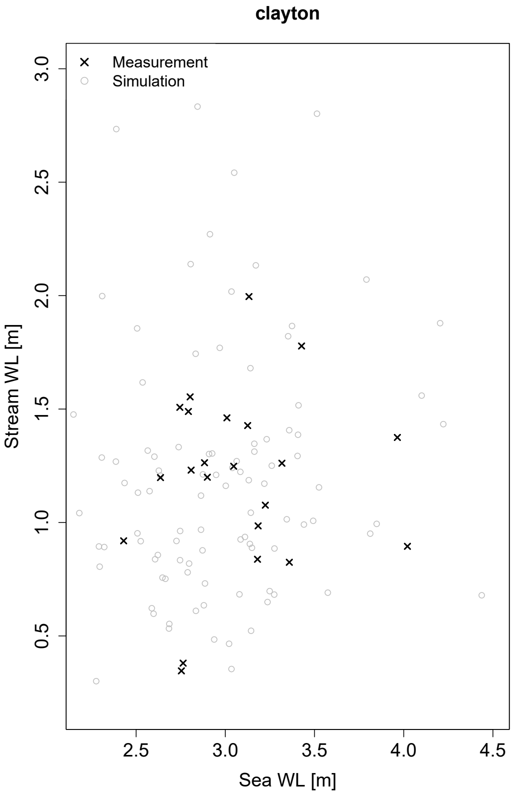


Figure 37 Scatter plot for Ribe 4 data

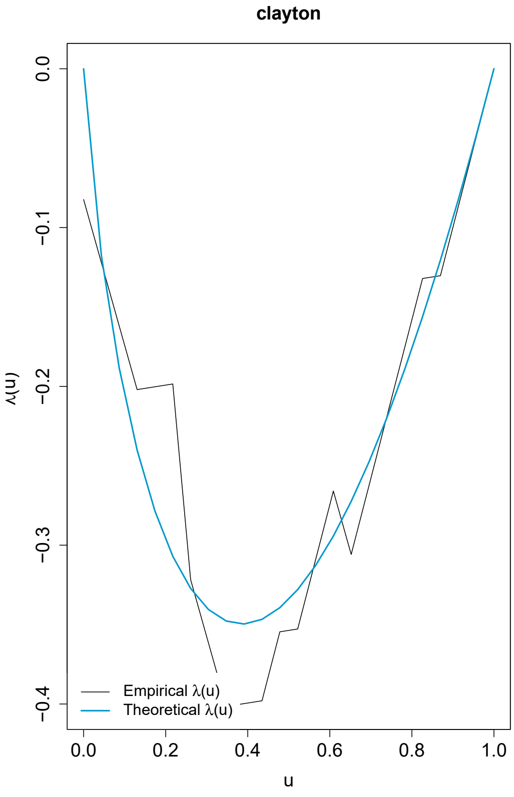


Figure 38 Lambda copula plot for Ribe 4

6.4.5 Copula plot

With 20 random pairs plotted on the 100 year return interval isoline (left). The 20 pair sample is shown in the table to the right.

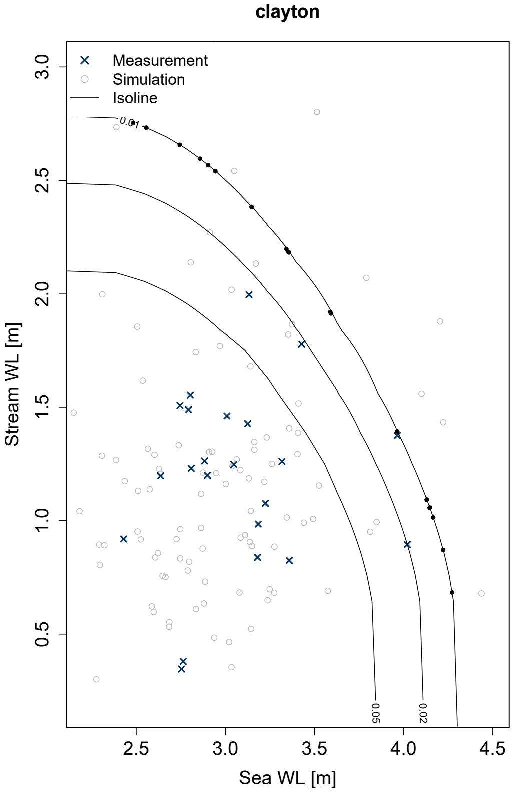


Figure 39 shows the copula plot from Ribe 4

Pair	Sea (m)	Stream (m)	Pair	Sea (m)	Stream (m)
1	4.13	1.09	11	2.56	2.73
2	2.48	2.75	12	2.94	2.54
3	4.22	0.87	13	3.59	1.92
4	3.35	2.19	14	2.86	2.60
5	2.74	2.66	15	4.15	1.06
6	4.27	0.68	16	4.17	1.01
7	3.34	2.20	17	4.15	1.06
8	3.96	1.39	18	3.36	2.18
9	4.13	1.09	19	2.90	2.57
10	3.15	2.38	20	3.59	1.91

Table 10 shows the 20 random pairs from isoline MT100 in the Ribe 4 copula plot

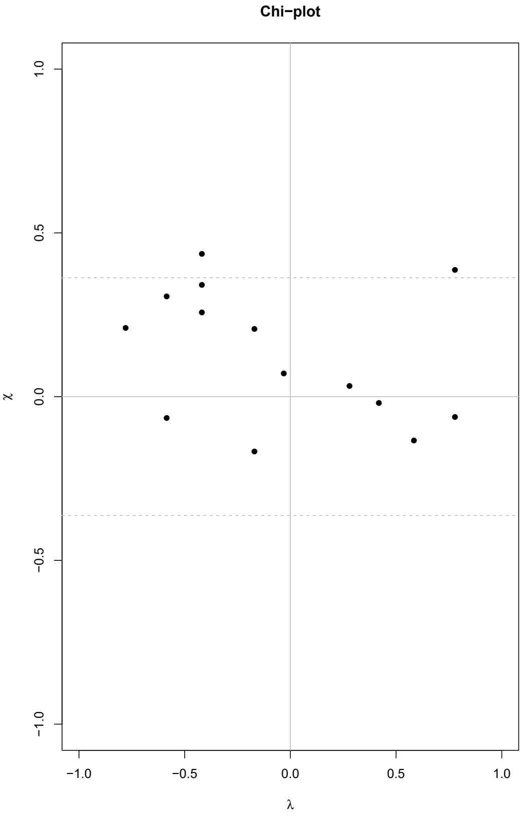
6.5 Ribe 5

6.5.1 Sample

1. value pr. year Date	Hydro. year	Sea (m)	Stream (m)	2. value pr. year Date	Hydro year	Sea (m)	Stream (m)
26-10-2005 07:00	2006	2.01	1.77	15-11-2005 01:00	2006	1.99	1.62
12-01-2007 07:00	2007	3.13	2.10	18-03-2007 15:00	2007	2.89	1.55
01-03-2008 19:00	2008	3.32	1.64	01-02-2008 09:00	2008	3.05	1.62
04-09-2009 13:00	2009	2.76	1.28	10-11-2008 11:00	2009	2.43	1.53
18-11-2009 15:00	2010	3.22	1.78	04-10-2009 03:00	2010	2.75	1.23
05-02-2011 04:00	2011	2.90	1.65	12-11-2010 19:00	2011	2.81	1.80
09-12-2011 12:00	2012	2.80	1.94	03-01-2012 22:00	2012	2.79	1.90
31-01-2013 03:00	2013	3.43	1.99	25-11-2012 23:00	2013	2.64	1.84
05-12-2013 16:00	2014	4.02	1.61	28-10-2013 16:00	2014	3.18	1.71

Table 11 shows the sample from Ribe 5

6.5.2 Chi- plot and Kendall's plot for dependence



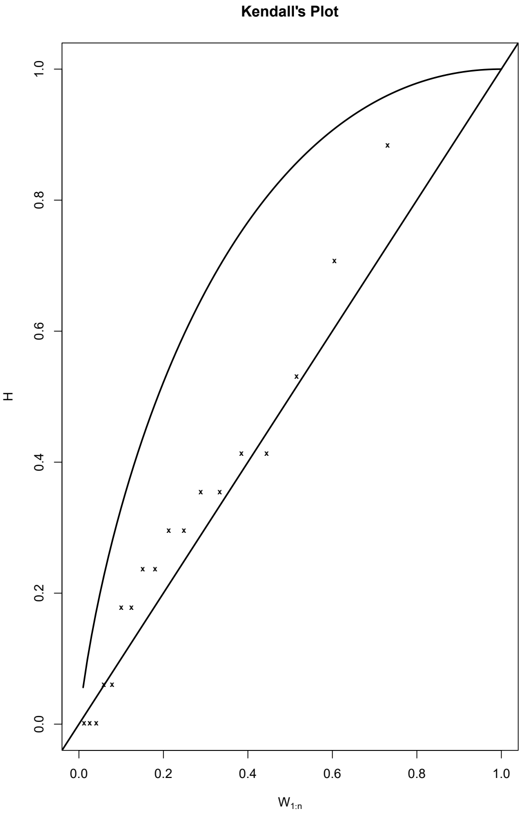
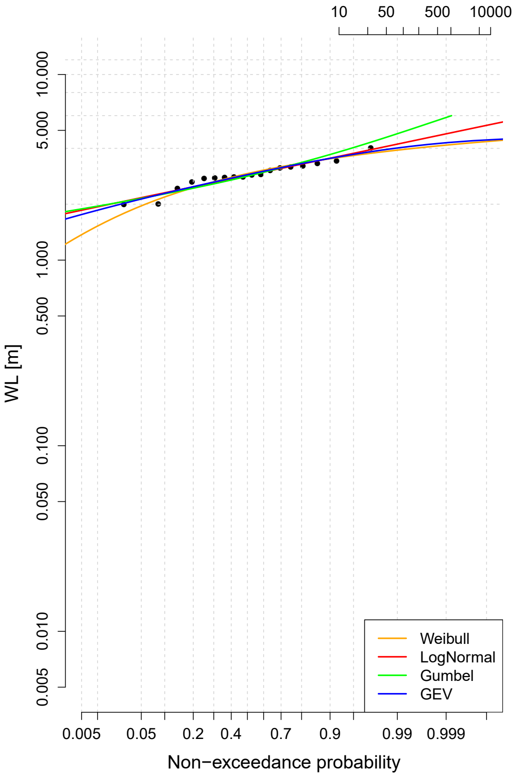


Figure 40 Ribe 5's Chi-plot

Figure 41 Ribe 5's Kendall's plot

6.5.3 Return time plots for Sea (left) and Stream (right)



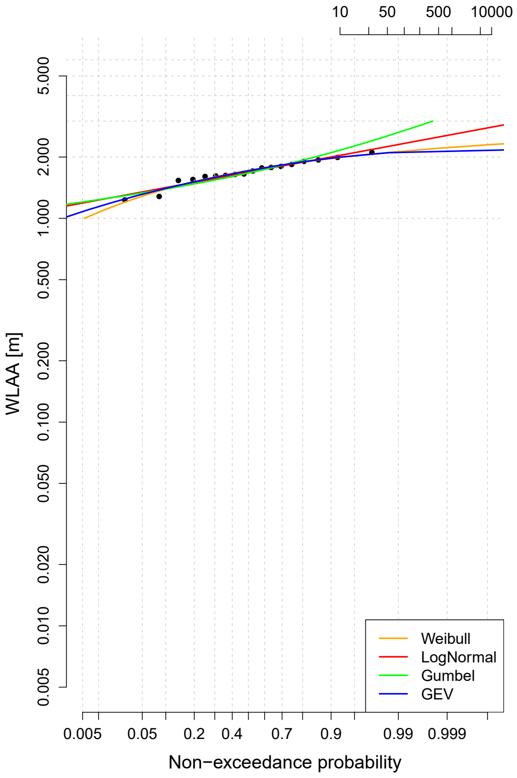
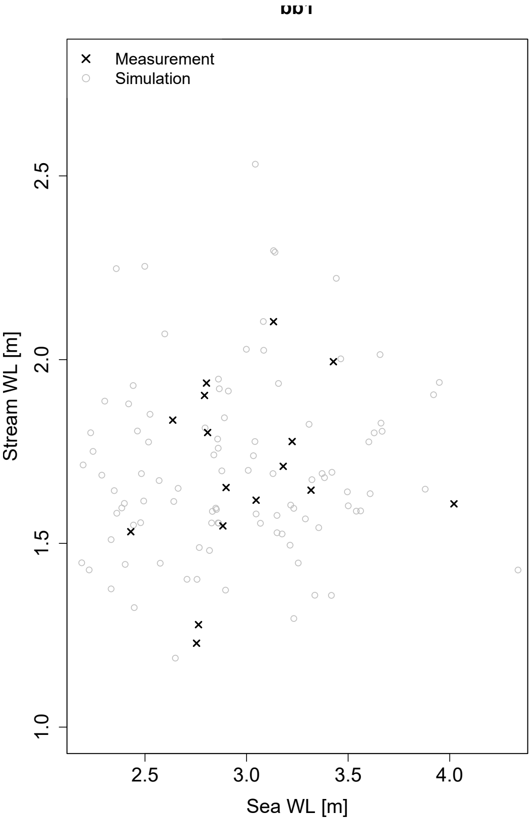


Figure 42 Return plot for Ribe sea (5)

Figure 43 Return plot for Ribe 5 (stream)

6.5.4 Copula fitting: Scatter- and lambda -plot



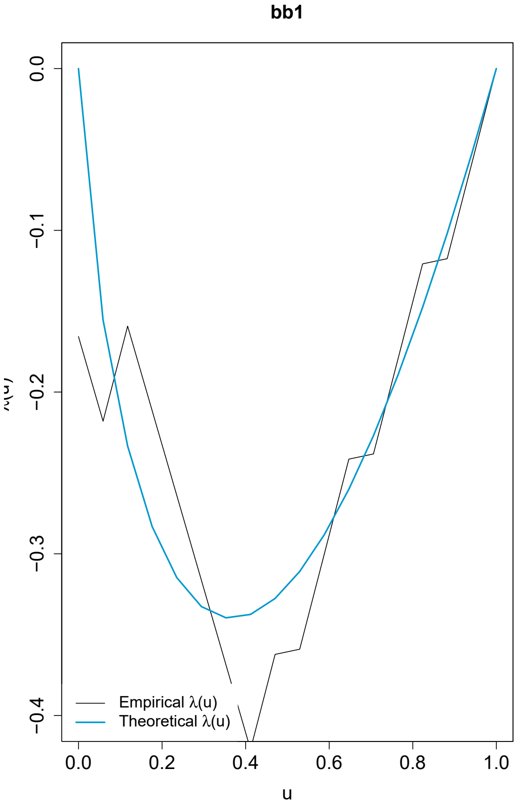


Figure 44 Scatter plot for Ribe 5 data

Figure 45 Lambda copula plot for Ribe 5

6.5.5 Copula plot

With 20 random pair plotted on the 100 year return interval isoline (left). The 20 pair sample is shown in the table to the right.

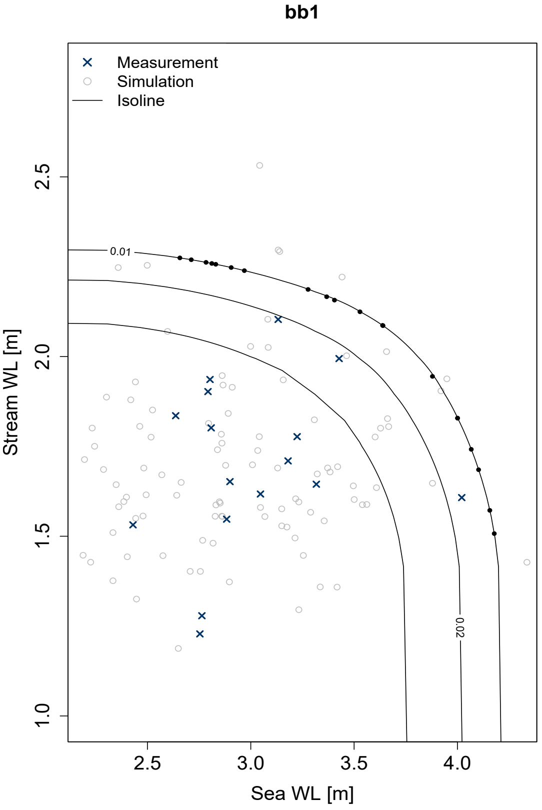


Figure 46 shows the copula plot from Ribe 5

Pair	Sea (m)	Stream (m)	Pair	Sea (m)	Stream (m)
1	3.64	2.09	11	2.83	2.26
2	2.66	2.27	12	2.97	2.24
3	4.00	1.83	13	4.00	1.83
4	3.37	2.17	14	4.10	1.69
5	3.64	2.09	15	4.16	1.57
6	2.71	2.27	16	3.41	2.16
7	3.53	2.12	17	4.18	1.51
8	3.28	2.19	18	2.81	2.26
9	3.88	1.94	19	2.91	2.25
10	4.07	1.74	20	2.78	2.26

Table 12 shows the 20 random pairs from isoline MT100 in the Ribe 5 copula plot

7. Discussion of data

7.1 Joint Probability Comparison between Stations

The joint probability results were presented through 20 data pairs (for the 100-yrs. return period) for each station. It is determined that the 20 pairs are representative of scenarios in the middle-to-high end of water levels in both sea and stream. The lower scenarios are of little importance since typically they do not lead to flooding of a significant magnitude. The scenarios to be included for each station should be event combinations i.e. scenario A. high water level in sea and medium level in the stream, or scenario B. medium level in sea and extreme level in the stream.

The description of sea and stream data below is mainly included as a way to evaluate examples of combinations.

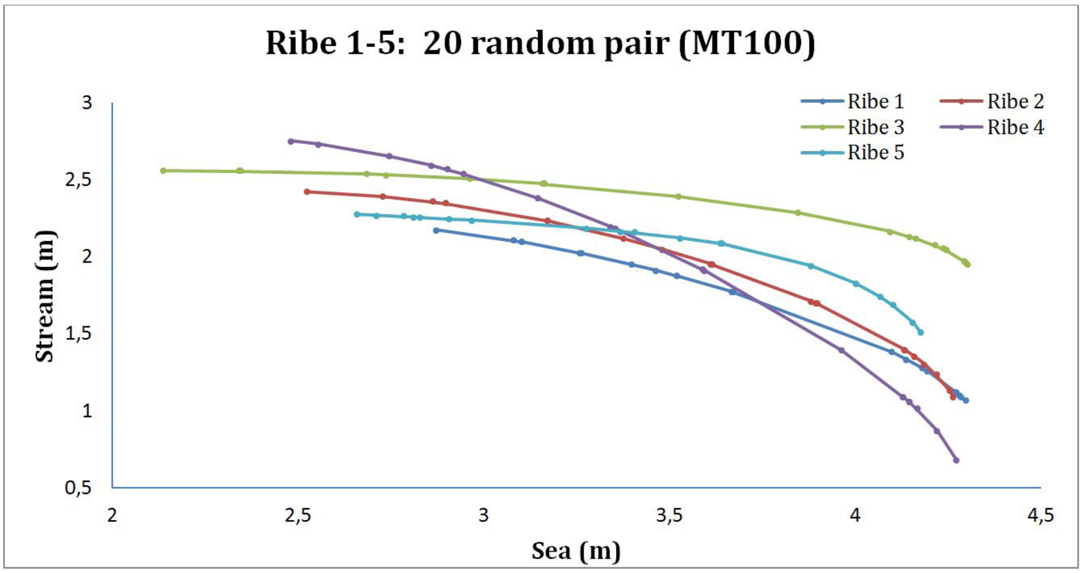


Figure 47 shows the randomly produced 20 data pair from the respective copula plot for each station. A tendency line is drawn for each station's data pairs.

7.1.1 Sea Data

In calculations, the two highest annual sea data sets constitute the first variable and stream data the second variable. Figure 47 above shows that at the lower end (left side of each curve) the minimum water level varies from 2.14 m DVR90 (Ribe 3) to 2.87 m DVR90 (Ribe 1). At the high end (right side of each curve) there is a 0.12 m variation between the minimum value of 4.18 m DVR90 (Ribe 5) and the maximum value of 4.30 m DVR90 (Ribe 3) (see map of Ribe in figure 2).

The largest interval in any station is in the data from Ribe 3, where the lowest sea level is 2.14 m DVR90 and the highest sea level is 4.30 m DVR90. This difference in lowest and highest water level is very significant but is not necessarily a controlling factor in regard to flood severity.

The smallest interval is found in Ribe 1 with a minimum of 2.87 DVR90 and a maximum of 4.28 DVR90, yielding a spread of 1.41 m.

This analysis only reflects the tendency of the 20 pairs but, keeping in mind the assumption made that the 20 pairs are representative of all data scenarios, the specific values are not as important as the tendencies.

Considering that Ribe is located approximately 5.5 km away from the sea, and that there is a sluice which closes with a 2-3 cm difference in water level between high sea and lower stream, it seems realistic that flooding in and around Ribe is caused mainly by the 'damming up' of water from the streams. This could indicate that the sea water level is less relevant, as long as the sluice is closed which is the case under most, if not all, storm surges conditions.

7.1.2 Stream Data

When focusing on the stream data intervals between the individual stations, it appears that in the lower end (left side of each curve) of stream data, there is a 0.57 m variation between the minimum (2.18 m DVR90, Ribe 1) and the maximum (2.75 m DVR90, Ribe 4). The higher end (right side of each curve) shows a minimum value of 0.68 m DVR90 and a maximum value of 1.97 m DVR90, yielding a total spread of 1.29 m. As the main driver in regard to potential floods around and/or in Ribe is the streams, a difference of 1.29 m seems noticeable. However, this variable is difficult to interpret as it is only reflecting the maximum value of the overlapping time series and not the full stream data series.

The largest difference between minimum and maximum of any stream water level is found in Ribe 4 with a minimum of 0.68 m DVR90 and a maximum of 2.75 m DVR90 and a total spread of 2.07 m. Likewise, the smallest interval is found in Ribe 3 with a minimum of 1.97 m DVR90, a maximum of 2.56 m DVR90, and a spread of 0.59 m.

The severity of a flood scenario typically depends on: Water level exceeding the natural protective structure (if present) threshold, the weather system previous to the storm surge i.e. wind, pressure, sediment - saturation ability, and precipitation etc., but also on whether there are constructions in place to handle large water inflows, the presence of vulnerable structures, the capacity of the environment to absorb or store the sudden increases in water inflow, tides etc.

7.2 Short time Series and Choice of Marginal Distribution

As previously mentioned, the sea time series in the statistics are limited by the stream time series resulting in no more than approximately 17 years of overlap. This affects the statistical return values shown in table 13.

Return time (yrs.)	HVS17 water level LogNormal (m)	Joint probability LogNormal (m) Fit (μ,σ): (1.08, 0.17)	Joint probability Gumbel (m) Fit (χ,α): (2.74, 0.40)
20	4.41	4.07 (25 yr. return)	4.22 (25 yr. return)
50	4.70	4.26	4.50
100	4.88	4.45	4.78

Table 13 Return probability and water level values from HVS17. Sampling method (pot threshold = 3.41m)
Distribution function: LogNormal

Taking the 100 years return value as an example, it is 4.88 m DVR90 for the HVS17 statistics based on 98 years of data (1919-2017). In comparison, the joint probability statistics based on approximately 13 years of data, the return value is reduced to 4.45 m DVR90 applying the same distribution – LogNormal, and 4.78 m DVR90 with the Gumbel distribution function. This rather significant difference is problematic when trying to create or improve a prognosis model, which is one of the purposes of this report. When using the Gumbel distribution function the return value is closer to the return value of HVS17, with a difference of just 10 cm, compared to the 43 cm from LogNormal.

For the joint probability statistics, it was decided to use the same distribution function for sea data as used in HVS17 for the purpose of comparison and quality check of data. In other words, 98 years of data is more reliable, and usually more tendency revealing, than 13 years of data. Although the Gumbel distribution is closer to HVS17 value, it is only based on 13 years of data. This is also backed by the fit of the distribution, shown for Lognormal as Fit (1.08, 0.17) and for Gumbel Fit (2.74, 0.40) (table 13): the lower the value, the better the fit.

7.3 Challenges Associated with Performing Joint Probability Calculations

Joint probability calculations are complicated. Many parameters may influence the outcome significantly such as: the sampling method, degree of dependence, marginal distribution, parameter fitting method etc. As every parameter seems to weigh heavily on the output, it is important to fully understand each parameter; however for most parameters appropriate methods to perform sensitivity analyses have been difficult to find..

7.3.1 Choosing Sampling Method

The chosen sampling method in this report is the AM method and it is based on a thorough evaluation of the different samples resulting from the two different sampling methods.

The choice of sampling is significant. This decision should depend on: how much data is available (form and length of time series), how complete the data series are (missing data etc.), the amount of extreme events present in data sets etc. There are two popular methods: Peak Over Threshold (POT) and Block Maxima (Annual Maxima or AM).

When working with extreme data, the POT method is a popular sampling method when the data series are long. This is, amongst other reasons, due to the possibility of controlling the definition of an extreme valuefor that particular region. However, when data is sparse and there are few or no actual extremes, it becomes difficult to evaluate whether a chosen threshold yields a misleading sample tendency.

The AM method is based on the highest annual value(s). With this method one risks a large variation in event sizes through the years of the data series; however it becomes very clear if a year's data should be regarded as an outlier and it can easily be removed before further calculations are performed.

In a minor sensitivity test performed alongside the development of the joint probability method, the AM method appears to be more useful when the amount of data is limited. For more information on the AM method, please refer to: On the block maxima method in extreme value theory by Ferreira, A. and De Haan, L. (2013).

One AM/year did not provide sufficient values to obtain a decent fit of the marginal distributions, whereas two AM/years improved the possibility of obtaining marginal distributions with acceptable fits. Values too low (outliers) compared to the total sample were removed in order not lower the defined extreme value level more than necessary.

The POT method was not used because of the variation between annual data. Some years had lower measurements that any other years and applying representative extreme level thresholds would exclude those years. In theory, this is perfect if focus is on "extreme" values (because one wants to remove values that are not considered extremes), however since   focus is on the combined effect of both stream and sea water levels, a high-to-extreme level in sea, these medium-high highs are also of interest. It is thus not always the worst storm surges in terms of extreme water levels that cause the most severe floods but a combined effect of several factors coinciding. This can be evaluated by not eliminating certain years due to low sea level measurements. In the context of the presented work, it is the sea data series that determine what stream data are recorded from the overlapping time interval.

The result of a sensitivity test between the BM and POT method showed that approximately 4 years of data was removed out of the 13 years of total data when using POT with a threshold of 2.73 m DVR90 (the lowest annual value of all available years. The actual extreme level of e.g. a 20-yr. return time for Ribe is 4.41 m DVR90); see appendix A. Since this report has its focus on the combined effect of high-to-extreme water level and because there is a general interest in using as many years of data as possible, the BM method is preferred.

It is recommended to test both methods and to evaluate their respective samples to the specific purpose.

7.3.2 Chi and Kendall's

A factor of great importance is the degree of dependence between the two variables. The Chi plot and the Kendall's plot are the visual outputs that need evaluation. The expert performing the calculations should be able to assess whether the dependence is sufficient to allow copula functions to be used or not. Finding information on evaluations of Chi- and Kendall's plots is not impossible but specific cases that relate to joint probability calculations on water level/flow studies have not beenfound in literature.

7.3.3 Tau and Theta

The value tau describes the correlation between the sea data and the stream data. As mentioned in section 4.5.1.1 tau varies between -1 and 1, with O being a no-correlation scenario. However, data are rarely uncorrelated with an exact zero value. This means that in performing the calculations, some experience is required in evaluating tau values of different variable combinations to avoid misinterpretation and thus the risk of creating an inaccurate picture of the statistical scenario.

In general, most of the calculations demand a certain level of experience, which can be obtained simply through observing and evaluating a large number of combination scenarios and their tau values.

Theta is a parameter that assists the assessment of the copula fit. The lower the value, the better the fit is. This, unlike the tau value, does not require much experience since the lowest value must be the best fit for the options given. However without experience this can, once more, become a difficult decision to make.

7.3.4 Choosing Marginal Distribution

The presented results do not include information on any more than the chosen distribution and a brief presentation of the remaining distribution types are discussed below. A distribution, such as the GEV distribution, is not discussed because it turns out to be too sensitive to values in the lower end (compared to the extreme values).

7.3.4.1 Sea

In the decision-making regarding which distribution function to apply to fit the sea data, the following was considered:

- The best fit by the lmom method

Below is shown a table with the lmom fit parameter for Ribe 1. The lower the value the better the fit is, and here it is obvious that the LogNormal distribution provides the best fit of the three.

lmom fit for Ribe Sea data	Arithmetic mean	Standard deviation
Weibull	7.03	3.18
LogNormal	1.08	0.17
Gumbel	2.74	0.40

Table 14

- The HVS17 statistics are based on a longer time series (approx. 98 years) and the best fit for this data is found with the LogNormal distribution, which s amongst others, is a reason for using the same distribution in the joint probability calculations.

Below are two tables showing a. the return values for Ribe sea for 25-, 50- and 100-yrs. return periods and the three distribution functions of interest and b. the return values from the LogNormal values from the HVS17 report.

If the joint probability values (left table) are related to the HVS17 values (right table) none of the distributions seem to fit well. If focusing on the 100 yr. return period, the Weibull distribution gives a value of 0.57 m lower than the value from HVS17. The LogNormal is once more lower than the HVS17 value by 0.14 m and the Gumbel distribution overestimates HVS17 value by 0.14 m.

Joint probability

Return probability	Weibull (m)	LogNormal (m)	Gumbel (m)
0.04 (25 yrs.)	4.03	4.18	4.25
0.02 (50 yrs.)	4.20	4.51	4.69
0.01 (100 yrs.)	4.31	4.74	5.02

Table 15

Return probability	Water level (m)
0.05	4.41
0.02	4.70
0.01	4.88

Table 16

It becomes clear that the choice of distribution depends on for what the values, or thresholds, should be used. Finding a fitting extreme water level protection threshold should not be underestimated – since it could result in flooding. Financially, building according to overestimated thresholds would increase costs but if the ultimate purpose is to reduce/eliminate the risk of flooding an underestimation of threshold is ineffective.

7.3.4.2 Streams

Choosing the right marginal distribution function for the stream data is to some extent more complicated than it is for the sea data. This is mainly due to the lack of comparable data. In the case of the sea data, statistics (HVS17) have been calculated based on longer and much more robust data than the data used for the joint probability calculations. This provides the opportunity to evaluate the short joint probability data series against the long robust HVS17 data series and thereby evaluate the choice of distribution. The same is not possible for stream data.

A consideration, in regard to stream data and choice of distribution, is provided in table 17.

Ribe 1 - Stream data

Return probability	Weibull (m)	LogNormal (m)	Gumbel (m)
0.04	1.76	1.84	1.86
0.02	1.86	2.04	2.10
0.01	1.93	2.18	2.28

Table 17 Stream Data: return values for a 25-, 50- and 100-years return period and the three distribution functions of interest.

Knowing that the LogNormal underestimated the water level by 14 cm in the sea compared to the HVS17 data, and assuming that the HVS17 data is more reliable based on the longer data series, it could be assumed that a similar relationship is found in the case of the LogNormal and the Gumbel distribution function used on the stream data.

Focusing on the LogNormal and Gumbel distributions, the 100-yrs return periods show a 10 cm difference. The consequence of choosing one distribution over the other does, in this case, not appear significant. However, the "true" picture of the streams maximum events are not necessarily included in the calculations since it is the sea variable that determines the stream data considered.

7.3.5 Choosing the Copula Distribution

Choosing the right copula distribution is similar to making the choice of the marginal distribution and thereby it becomes much easier, even if a performer's experience with copulas is limited, and the process appears to be more natural – not so risky and uncertain. As mentioned in the methods section (section 4.5.5 and 4.5.1.5), there are two ways of evaluating the fit of a distribution function; the graphic (visual) and the mathematical way. The visual inspections are based on several graphs and the relation between theoretical and empirical data. The mathematical "fit" is assessed by the theta value (see table 2 in section 6)

7.4 Choosing Combinations of Water Level in Sea and Stream in regard to Flooding caused by Combined Events

Depending on which of the two sources that appears to be the dominating flooding source, one may choose combinations which will be biased towards high water level in the stream or high water level in the sea.

In this case, we know what the critical levels are for the respective sources but what can be done where the critical thresholds are unknown?

One method is to focus on the tendency in the actual data. Presented below is a plot showing the data from a copula plot and a linear tendency line based on the actual data points and 50 random pairs on the MT50 isoline (circles). The general idea is that by looking at the tendency line, the most probable combinations will be depicted around the tendency line. Defining an interval on each side of the tendency line provides more combinations, and it is expected that these are the combinations occurring with higher frequency.

One problem with this selection of combinations method is that the joint probability method itself already weighs towards the high sea levels and not necessarily high stream levels. This means that when the interval of combinations is chosen there is a risk that it will show more combinations leaning towards high sea level – medium/low stream level. This could very well be misleading but can only be evaluated by analyzing the effect of using stream data as first variable (VAR1). Read more about this below.

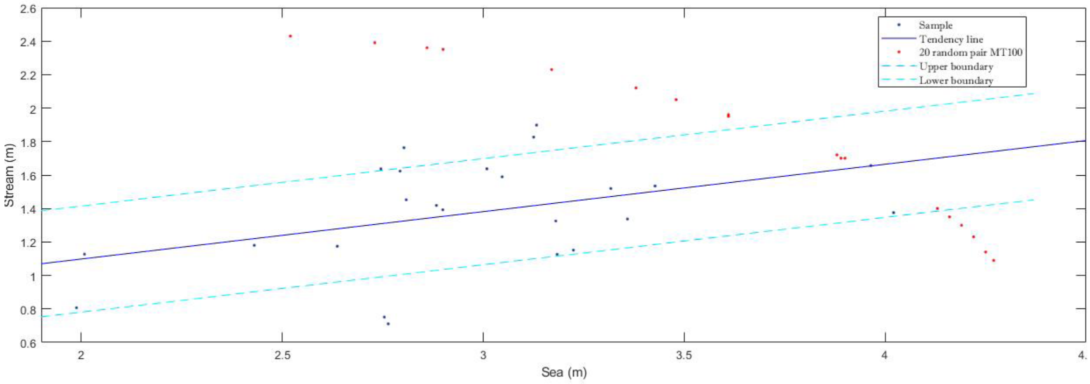


Figure 48 shows a plot of a sample with a trend line and its lower and upper boundaries (2σ). It is suggested to focus on combinations within the boundary interval.

8. Conclusion and Suggested further Work

The method presented was recently introduced to the DCA. It proved rather difficult to use Ribe as a pilot test site because of the the sluice. The sluice limits the actual contact between stream and sea water during a storm surge. However, the damming effect of water in the streams near the outlet (by the sluice) is still occurring due to the lack of outflow. Interpretation of the results are difficult but with the next step (MIKE modeling using the joint probability results) the picture is expected to become lucid.

Several minor analyses were performed to identify and determine which method would appear to be best for: sampling data, choosing marginal distribution function, choosing copula and finally calculating the joint probability of high-to-extreme water level situations occurring simultaneously in the area around and inside Ribe.

In regard to Future Aspects or Research the Following is recommended:

More and better Data

The longer the data series, the more possibilities present themselves. The main problem with the short time series is the lack of actual extreme values. Keeping in mind that water levels of high-to-extreme is of interest in this project, and not only the extremes, it becomes easier to accept the values that can be worked with, but they are actually only representing the "high" end, not extremes.

Analyze the effect of the sluice

The sluice closes when the seas water level is 2-3 cm higher than the streams water level but how this affects the stream remains to be investigated and a clarification is therefore recommended.

What Behavior or Tendency is to be observed when Streams become first Variable?

Another really interesting aspect is to look at the same data but with the streams as first variable instead of the sea. The sea is affecting the stream and not the other way around, however to ensure that the highest stream levels are presented, stream data would need to be the first variable. It would be interesting to investigate whether the return times and values would change and if so, how?

Analysis of Weather Systems resulting in High-to-extreme Water Level

It would be interesting to look more closely at weather systems and how they affect the statistics. Ribe's stream systems are definitely under stress when western wind systems push the water front inland (even with the sluice because of the stagnation of stream water).

Use more Sources (primarily Precipitation and Groundwater) to obtain a more complete Picture

Precipitation is expected to be a major indirect and or direct source in flooding from streams, but it is not included in the analysis. The same goes for groundwater. For a better picture of the whole system these two parameters should be analyzed and included. Precipitation can easily be linked to the weather analysis so it is recommended to do these analyses simultaneously.

Climate Changes and Climate Contribution

Calculations with climate contributions for the given return periods should be performed and the result should be analyzed.

The Danish Meteorological Institute (DMI) will provide new climate contribution values for certain return periods, which will be included in later analyses.

9. References

Dangendorf, S., Marcos, M., Wöppelmann, C., Conrad, C., Frederikse, T. and Riccardo, R. (2017). Reassessment of 20th century global mean sea level rise. PNAS, Vol. 114 no. 23.

Knudsen, P., Khan, S. A., Engsager, K. S., & Sorensen, C. (2016). An uplift model for Denmark – and work ahead. Frontiers in Marine Science, 3, [69; Supplementary Material]. DOI: 10.3389/fmars.2016.00069

Maria Ivette Gomes, Armelle Guillou. Extreme Value Theory and Statistics of Univariate Extremes: A Review. International Statistical Review, Wiley, 2015, 83 (2), <10.1111/insr.12058>. <hal-01311707>

Marchi, V., Rojas, F. and Louzada, F. (2012). The Chi-plot and its asymptotic confidence interval for analyzing bivariate dependence: An application to the Average Intelligence and Atheism Rates across Nations Data. Journal of Data Science 10 (2012), p. 711-722.

Vogel, R. M., and Wilson, I. (1996). Probability distribution of annual maximum, mean, and minimum stream flows in the United States. Journal of Hydrologic Engineering 1, 69-76.

https://ascelibrary.org/doi/10.1061/%28ASCE%291084-0699%281996%291%3A2%2869%29

Dombry, C. (2013). Maximum likelihood estimators for the extreme value index based on the block maxima method. 18p. <hal-00780279>

Nelsen, Roger B. (1997). Dependence and order in families of Archimedean copulas. Journal of Multivariate Analysis 60, p. 111-122. Article no. MV961646

Ferreira, A. and De Haan, L. (2013) On the block maxima method in extreme value theory: PWM estimators. The Annals of Statistics 43(1). DOL: 10.1214/14-AOS1280

Kystdirektoratet
Højbovej 1
7620 Lemvig
www.kyst.dk