universität
wien

MASTERARBEIT / MASTER’S THESIS

Titel der Masterarbeit / Title of the Master‘s Thesis

„Risk Capital Reserves for Flood Catastrophes in National and European Context“

verfasst von / submitted by

Andreas Wittmann BSc.

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc.)

Wien, 2017 / Vienna 2017

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 920
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Quantitative Economics, Management and Finance UG2002
Betreut von / Supervisor:	o. Univ.-Prof. Mag. Georg Pflug
Mitbetreut von / Co-Supervisor:	

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent years the damages caused by floods in Europe and the rest of the world have been increasing. While some of the increased damages can be explained with higher vulnerability it also seems that both frequency and severity of flooding events increase across the globe. This increase can at least to some extent be traced back to climate change.

The impact of climate change is especially worrying as any change in the weather pattern affects all bodies of water in a larger area at the same time. Therefore climate change increases the probability that many rivers burst their banks simultaneously and cause significant damages in a wide region. With this in mind the damage associated to the individual rivers cannot be viewed as independent, which puts a limit to the extent that the risk posed by flood damages can be diversified by insurance companies.

A sufficient level of insurance is vital for the economy and it has been shown that with adequate insurance the impact of a catastrophic event can be mitigated to a large extent, as provided sufficient funds are available in the aftermath of catastrophes an increase in innovation and investment can be observed.[20]

The predicted changes to the climate raise the question whether with current contracts sufficient insurance will remain available to the public in the coming years, as those changes will lead to increased insurance premia in some regions and possibly make insurance completely unavailable in certain areas as insurance companies withdraw from no longer profitable markets.
It is therefore important to find estimations not only for the damages caused by the individual river beds but also for the sum of total damages in an area.

With this motivation in mind, in this work we try to find distributions for the total flood damages in Europe to enable us to estimate the risk posed by river floods. For this task we utilize R-Vine-copulas.

The structure of the work will be the following. In chapter 2 we will introduce the main theoretical concepts of this work. The chapter will not be exhaustive and additional methods and concepts will be introduced as they are needed. Chapter 3 will discuss the available data and the decisions made by the author originating from it. The chapter 4 will then showcase a small R-Vine model resulting from the data and chapter 5 will give the forecasts for two key years, 2020 and 2085 derived with the methods of this work.
We decided to collect additional tables in appendix A and some proofs in B.

Chapter 2

Important concepts and definitions

In this chapter we will present the most important concepts and methods we apply throughout this work. The goal is to explain the used methods to the reader and provide some insights into these topics.

2.1 Kendall's τ

Definition 1. ***Concordance***
Let (X, Y) be a vector of two random variables, and let $s_0 = (x_0, y_0)$, $s_1 = (x_1, y_1)$ be two samples drawn from (X, Y)).
s_0 and s_1 are called concordant if

$$(x_0 - x_1)(y_0 - y_1) > 0$$

and they are called discordant if

$$(x_0 - x_1)(y_0 - y_1) < 0.$$

Pairs which are neither are called draws.

With the help of definition 1 M.G. Kendall defined the following correlation coefficient in 1938.[12].

Definition 2. ***Kendall's*** τ
Let $(X, Y) = (x_i, y_i), \quad i = 1, ..., n$ be a vector of n observations coming from two random variables. $\tau_{X,Y}$ is defined as

$$\tau_{X,Y} := \frac{|\{concordand\ pairs\}| - |\{discordant\ pairs\}|}{|\{all\ pairs\}|}.$$

In the above definition draws which occur if one of the random variables takes the same value in two or more observations are not part of the numerator but are included in the denominator, so that for n observations the denominator in the above formula will always take the value $\frac{n^2-n}{2}$.

Kendall's τ takes values ranging from -1 to 1. A high absolute value of Kendall's τ is an indicator for a high dependence in the random variables, while for independent random variables Kendall's τ should converge to zero if the number of observations increases to infinity. As Kendall's τ is non-parametric it can be applied to an empirical sample without making assumptions about the distribution of the random variables first.

Kendall's τ is closely related to the Bubble-Sort-Distance d_{Bubble}.
The Bubble-Sort-Distance between two orderings A, B of an variable is defined as the number of switches a Bubble-Sort algorithm needs to make in order to perform the transformation of one ordering into the other.
A Bubble-Sort algorithm goes through the observation of variable in A and B and switches the location of the observations in B in every consecutive, discordant pair $(a_i, b_i), (a_{i+1}, b_{i+1})$. After reaching the final observation in A it starts again with the first variable until the variables in B are in the same order as the variables in A.
This way the Bubble-Sort algorithm switches every discordant pair in (A, B) exactly once so that, in the absence of draws, the relationship

$$\tau_{A,B} = \frac{\binom{n}{2} - 2d_{Bubble}(A, B)}{\binom{n}{2}}$$

holds.

The definition of the Kendall's τ-distance d_τ and our own adaptation, which we call the information distance $\hat{d_\tau}$, is given in the following.

Definition 3. $\mathbf{d}_\tau$ *and* $\hat{\mathbf{d}_\tau}$

$$\begin{aligned} d_\tau(X, Y) &:= 1 - \tau_{X,Y} \\ \hat{d_\tau}(X, Y) &:= 1 - |\tau_{X,Y}| \end{aligned}$$

d_τ can take values in the interval $[0, 2]$, $\hat{d_\tau}$ takes values in $[0, 1]$.

It should be noted that $d_\tau(X, Y)$ are only metrics if one considers equivalence classes.

$$\hat{d_\tau}(X, Y) := min\{d_\tau(X, Y), d_\tau(X, \overleftarrow{Y})\}$$

where for $Y = (y_1, y_2, ..., y_n)$ we define $\overleftarrow{Y}$ as $\overleftarrow{Y} := (y_n, y_{n-1}, ..., y_1)$.

The proof that d_τ does satisfy the four requirements of a metric is rather straightforward and can be found in several sources, so that we decide to not repeat it in this work.
To the authors knowledge our information distance has not been introduced yet, therefore a proof that $\hat{d_\tau}$ does fulfil all requirements of a metric is given in B.
We find $\hat{d_\tau}$ advantageous compared to d_τ for our work, as it is only affected by the magnitude of the correlation but not by the sign of the correlation coefficient. This is beneficial to our work as we are mostly interested in the dependency of our random variables, and only to lesser extend whether this dependency increases (positive correlation) or decreases (negative correlation) the total risk.

2.2 Copulas

Definition

Most definitions for the copulas in this section are taken from the book *An introduction to copulas* by R.B. Nelson.[16]

Definition 4. ***Copula***
A n-dimensional copula is a function C from $[0,1]^n \to [0,1]$ with the properties that

- *For every $\mathbf{u} \in I^n$ if at least one coordinate of $\mathbf{u}$ is zero*
$$C(\mathbf{u}) = 0.$$
- *If all coordinates of $\mathbf{u}$ except u_k are equal to one*
$$C(\mathbf{u}) = u_k.$$
- *For all $\mathbf{a}$ and $\mathbf{b} \in I^n$ such that $a_k \leq b_k \quad k = 1, ..n$*
$$V_C([\mathbf{a}, \mathbf{b}]) \geq 0.$$

V_C is the volume given by C for the n-orthotope[1] that is parallel to I^n and defined by the lower point $\mathbf{a}$ and higher point $\mathbf{b}$,

$$V_C([\mathbf{a}, \mathbf{b}]) := \sum_{\mathbf{d} \in \prod_{k=1}^{n} \{a_k, b_k\}} (-1)^{N(\mathbf{d})} C(\mathbf{d}), \qquad N(\mathbf{d}) := |\{k | d_k = a_k\}|.$$

[1] The n-dimensional generalisation to the two-dimensional rectangle and the three-dimensional cuboid

In 1959 A. Sklar showed that every multivariate distribution function can be expressed though the marginal distribution functions and a coupling function, the copula.[18]

Theorem 1 (Sklar)**.** *Let H be a joint distribution function with marginal distribution functions F and G. Then there exists a copula C such that for all x,y*

$$H(x,y) = C(F(x), G(y)).$$

C is unique if both F and G are continuous. Otherwise it is uniquely determined on the the range of the marginal distribution functions F and G.

The proof for this theorem in the case of continuous margins is rather straightforward and is given in B. Proof for cases with non-continuous margins are more involved and can be found in the book by Nelson or papers by Sklar.

Theorem 2. ***Fréchet-Hoeffing bounds***
Define

- $M(u,v) := min\{u,v\}$,
- $W(u,v) := max(u+v-1,0)$.

Then for every two-dimensional copula $C(u,v)$ the following inequality holds

$$W(u,v) \leq C(u,v) \leq M(u,v)$$

The upper Fréchet-Hoeffing bound M(u,v) describes co-monotonicity in the random variables, the lower Fréchet-Hoeffing bound W(u,v) anti-monotonicity In two dimensions both bounds are copulas themselves, for more than two dimensions only the upper bound remains a copula, while the lower bound looses the properties of a copula.

The field of copulas is rich and copula families to cover many different dependency structures exist. However since we only have a relatively small number of observations for each of our random variables available we mostly restrict ourselves to one-parametric copulas in two dimensions for the purposes of this work. In the following we will introduce the copula families used throughout this work by giving their definitions and showing some examples of random variables linked by them.

All of the following copulas are defined in the two-dimensional case that we need, though higher dimensional versions do exist.

Archimedian Copulas

Definition 5. ***Archimedian Copula***
*Let f(t) be a strictly decreasing, continuous and convex function that maps the unit interval to the interval (0,*inf*), such that f(1)=0. Define*

$$f^{-1}(t) := \begin{cases} f^{-1}(t) & 0 \leq t \leq f(0) \\ 0 & f(0) \leq t \leq \inf \end{cases}$$

then

$$C(u,v) := f^{-1}(f(u) + f(v))$$

is an Archimedian copula.

Theorem 3. *If C is an Archimedian copula, then C is*

- *symmetric,* $C(u,v) = C(v,u)$*,*
- *associative* $C(C(u,v),w) = C(u,C(v,w))$*.*

These two properties follow directly from the definition of an Archimedian copula and can be easily verified.

The generator functions and admissible parameter values for the copulas that will be discussed in the following pages can be looked up in table A.1.

The first Archimedian copula presented is the Product copula.

Definition 6. ***Product copula***

$$C_{Product}(u,v) = uv$$

The product copula is the simplest Archimedian copula and represents the case when the two coupled distributions are independent from each other.

Clayton family

Definition 7. ***Clayton copula***

$$C_{Clayton,\theta}(u,v) = max([u^{-\theta} + v^{-\theta} - 1]^{-\frac{1}{\theta}}, 0), \qquad \theta \in [-1, \infty) \setminus \{0\}$$

For $\theta \to$ inf the Clayton family converges to M(u,v), for $\theta = -1$ the Clayton copula is identical to W and for $\theta \to 0$ the family converges to the Product copula.

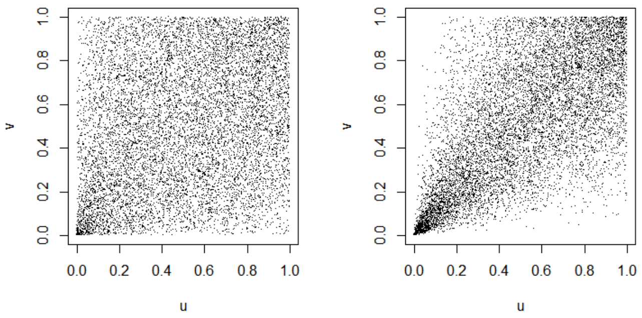

Figure 2.1: Variables linked by different Clayton copulas with the parameters $\theta = 0.5$ and $\theta = 2$.

Joe Family

Definition 8. ***Joe copula***

$$C_{Joe,\theta}(u, v) = 1 - [(1-u)^{\theta} + (1-v)^{\theta} - ((1-u)(1-v))^{\theta}]^{\frac{1}{\theta}} \qquad \theta \in [1, \infty)$$

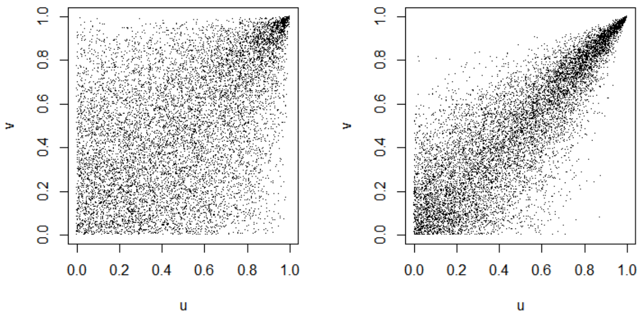

Figure 2.2: Variables linked by two different Joe copulas with parameters $\theta = 2$ and $\theta = 5$.

For the Joe copula family dependence increases with the parameter and it also converges to the upper Fréchet-Hoeffing bound M(u,v) for $\theta \to$ inf. However in contrast to the Clayton family the Joe-family does not reach the lower bound W(u,v). For the lowest admissible parameter $\theta = 1$ the Joe copula becomes to the Product copula.

Gumbel family

Definition 9. ***Gumbel Copula***

$$C_{Gumbel,\theta}(u,v) = exp(1 - [(1 - ln(u))^{\theta} + (1 - ln(v))^{\theta} - 1]^{\frac{1}{\theta}}) \qquad \theta \in [1, \infty)$$

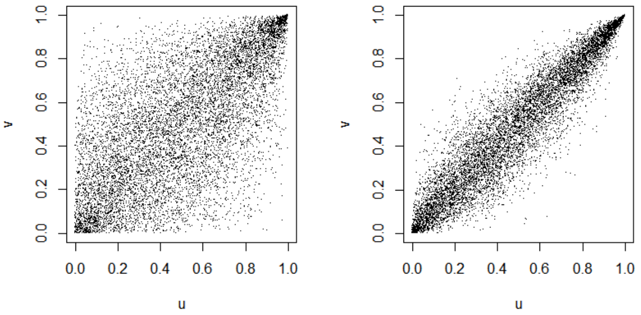

Figure 2.3: Variables linked by members of the Gumbel family with parameters $\theta = 2$ and $\theta = 5$.

The Gumbel family shows the similar behaviour as the Joe family, for low parameter values it is also close to the Product copula, while for $\theta \to \inf$ it converges to the upper Fréchet-Hoeffing bound.

Frank family

Definition 10. ***Frank Copula***

$$C_{Frank,\theta}(u,v) = -\frac{1}{\theta} ln(1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1}) \qquad \theta \in (-\infty, \infty) \setminus \{0\}$$

The Frank family also reaches the upper Fréchet-Hoeffing bound for $\theta \to \inf$, however in contrast to the last two examples and similar to our first example the Clayton family, the Frank family converges to the lower Fréchet-Hoeffing bound when $\theta \to -\inf$. For $\theta \to 0$ the Frank family converges again to the Product copula, which exhibits no dependence in the data.

Elliptical Copulas

Since we restrict ourselves in this work to one parametric copulas we only consider one member from the range of Elliptical copulas, the Gaussian or Normal copula.

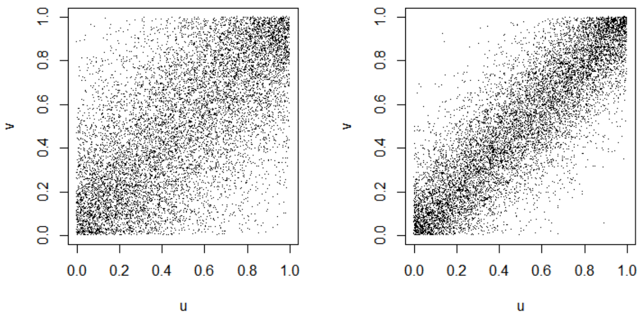

Figure 2.4: Variables linked by members of the Frank family with parameters $\theta = 2$ and $\theta = 5$.

Definition 11. ***Gaussian Copula***

$$C_{Gauß,\rho}(u, v) = \Phi_2(\Phi^{-1}(u), \Phi^{-1}(v), \rho_{u,v})$$

Φ_2 is the cumulative distribution function of a bivariate normal distribution, with Pearson correlation coefficient $\rho_{u,v}$ and Φ the cumulative distribution function of the standard normal distribution.

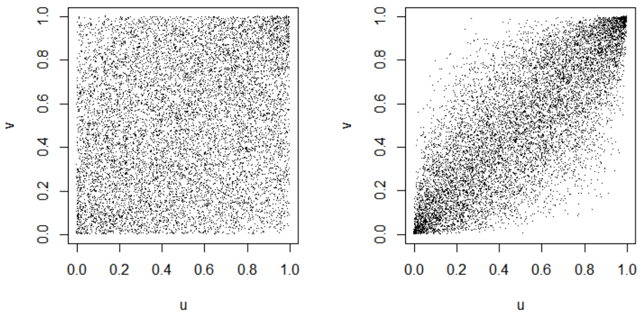

Figure 2.5: Variables linked by Gaussian copulas with $\rho = 0.2$ and $\rho = 0.8$.

The Gaussian family is able to reach both the lower Fréchet-Hoeffing bound, for $\Phi = -1$, and the upper Fréchet-Hoeffing bound for $\Phi = 1$ while for $\Phi = 0$ a Gaussian copula is identical to the Product copula.

Rotations

In order to alleviate the limitations we cause by restricting ourselves to this small selection of all possible copulas we also include rotations for the presented copulas. This allows us to better capture negative dependencies, as so far only three copulas are able to capture negative correlation. As a bivariate copula is defined on the unit square and all the copulas we introduced are symmetric with respect to the first median, only rotations by 90, 180 and 270 degrees can be considered. As all our copulas are symmetric regarding the main diagonal these rotations correspond to the base copula being flipped horizontally (270 degrees), vertically (90 degrees) or both (180 degrees). We denote the density functions of the resulting copulas as C^{90}, C^{180} and C^{270} respectively. The C^{180} version of a copula is often referred to as the survival version of the underlying copula C. The formulas for C^{90}, C^{180} and C^{270} can be derived from the underlying C in the following ways.

$$
\begin{aligned}
C^{90}(u,v) &= v - C(v, 1-u) \\
C^{180}(u,v) &= u + v - 1 + C(1-u, 1-v) \\
C^{270}(u,v) &= u - C(1-v, u)
\end{aligned}
$$

2.3 Tail-dependence

We are most interested in events when many river´s water discharge are simultaneously high and whether an extreme level in one river´s discharge contains information about the amount of water discharged by the other rivers in the same region.

In order to study the dependence conditional on at least one of our random variables taking either an extremely low or an extremely high value, that is to say the cumulative distribution function of the random variable is close to zero or respectively to one, we use the concept of tail-dependence. In *Quantitative Risk Management: Concepts, Techniques, and Tools* [6] the upper and lower tail dependence coefficients are defined in the following way.

Definition 12. ***Tail-dependence***
The tail-dependency of two random variables X, Y is defined as

$$\lambda_l = \lim_{q \to 0} P(F_Y(Y) \leq q | F_X(X) \leq q)$$

for the lower tail and

$$\lambda_u = \lim_{q \to 1} P(F_Y(Y) > q | F_X(X) > q)$$

for the upper tail, provided that these limits exists.

X and Y exhibit upper|lower tail-dependence or extremal dependence in the upper|lower tail if $\lambda_u|\lambda_u$ is different from zero. They are asymptotically independent in the upper|lower tail if $\lambda_u|\lambda_u$ is equal to zero.

Kendall's *tau* and the two tail-dependence coefficients λ_l and λ_u for two random variables X, Y can be expressed in terms of the copula $C(u, v) = C(F(x), G(y))$ linking the marginal distributions F and G of the random variables using the formulas below.

Theorem 4.

$$\begin{aligned} \tau &= 4 \iint_{[0,1]^2} C(u,v) dC(u,v) - 1 \\ \lambda_l &= \lim_{u \to 0} \frac{C(u,u)}{u} \\ \lambda_u &= \lim_{u \to 1} \frac{1 - 2u + C(u,u)}{1-u} \end{aligned}$$

The values for the Kendall's τ, λ_l and λ_u of the copulas used in this work are given in table A.2.

2.4 Pair-Copula-Construction, and Vines

Even though for most of the presented copulas higher dimensional versions do exist, we restricted all our definitions to the two-dimensional case so far.

Our reason for this restriction to the two-dimensional case is that even though copulas do present a powerful tool to model the dependency between random variables and Sklar´s theorem guaranties the existence of a copula linking the marginal distributions for an arbitrary number of dimensions, the formula and parameters for this copula is in most cases impossible to estimate from empirical data.
This is especially true when different subsets of the random variables exhibit different types and strength of dependency, like some subsets exhibiting upper tail dependence while other subsets show lower tail dependence.

To circumvent these drawbacks of copulas we use pair-copula-construction as it is presented by Bedford .[3] The pair-copula-construction provides a way to calculate the joint density function of any number of random variables by first factorizing the density into a product of conditional densities for the random variables

$$f(x_1, ...x_n) = f_1(x_1) * f_{2|1}(x_2|x_1) * ... * f_{n|1,2,...,n-1}(x_n|x_1, x_2, ...x_{n-1}).$$

The conditional densities appearing in the equation above are calculated in the following way.

$$\begin{aligned} f_{2|1}(x_2|x_1) &= c_{1,2}(F_1(x_1), F_2(x_2))f_2(x_2) \\ f_{3|1,2}(x_3|x_1, x_2) &= \frac{f_{2,3|1}(x_2, x_3|x_1)}{f_{2|1}(x_2|x_1)} \\ &= \frac{c_{23|1}(F_{2|1}(x_2|x_1), F_{3|1}(x_3|x_1))f_{2|1}(x_2|x_1)f_{3|1}(x_3|x_1)}{f_{2|1}(x_2|x_1)} \\ &= c_{23|1}(F_{2|1}(x_2|x_1), F_{3|1}(x_3|x_1))f_{3|1}(x_3|x_1) \\ &= c_{23|1}(F_{2|1}(x_2|x_1), F_{3|1}(x_3|x_1))c_{1,3}(F_1(x_1), F_3(x_3))f_3(x_3) \\ ... \end{aligned}$$

In this equations a small letter c stand for the density function of a copula C and is defined as

$$c(u, v) := \frac{\partial C(u, v)}{\partial u \partial v}.$$

The conditional cumulative distribution functions $F_{x|\mathbf{v}}$ necessary for the factorisation in the pair-copula-construction can be calculated as the derivative of the distribution function of a copula C . Joe.[10]

$$F_{x_i|\mathbf{v}}(x_i|\mathbf{v}) = \frac{\partial C_{x_i,x_j|\mathbf{v_{-j}}}(F(x_i|\mathbf{v_{-j}}), F(x_j|\mathbf{v_{-j}}))}{\partial F(x_j|\mathbf{v_{-j}})},$$

where $\mathbf{v_{-j}}$ is the vector $\mathbf{v}$ without the jth element. If $\mathbf{v}$ only contained one element the conditioning set $\mathbf{v_{-j}}$ is empty so that the unconditioned cumulative distribution function $F(x_i)$ can be used in the above formula to derive the conditional distribution function.

With the help of the pair-copula-construction the joint distribution of a high number of random variables can be derived. The pair-copula-construction

by itself however does not make a suggestion in which order the joint density should be factorized to achieve a good fitting model for a given dataset.
As estimation errors are unavoidable when working with empirical data and affect all future estimations that are conditioned on the copula fitted early it is reasonable to use the most informative, in the meaning that they showcase a high absolute value of correlation, pairs of variables in the dataset as early as possible in order to avoid loosing information due to estimation errors.

One way to determine the order in which copulas should be fitted to pairs of variables in the dataset is given by an R(regular)-vine. An R-vine is a sequence of nested trees fulfilling the requirements in the following definition.

Definition 13. ***R-Vine***
A sequence of trees $T_1, T_2, ...T_n$ is called a regular vine if it satisfies

- *$Edges(T_i) = Nodes(T_{i+1}) \quad 1 \leq i \leq n-1$.*
- ***Proximity condition** If two nodes in T_{i+1} are connected by an edge, the corresponding edges in T_i must share a common node.*

Definition 14. ***Constraint -,Conditioning - and Conditioned set***
The edges of a R-Vine are defined by the three following sets.

- *The constraint set of an edge e, $e = (v_i, v_j)$ in tree T_i is the union of the constraint sets of the edges in T_{i-1} that correspond to the endpoints of the edge e in T_i. The constraint sets for edges in T_1 consist of the two nodes at the ends of the edge.*
- *The conditioning D_e set of an edge is the intersection of the constraint sets of it´s two nodes.*
- *The conditioned set $C_e = (j(e), k(e))$ of an edge is the symmetric difference of the constraint sets of it´s two nodes.*

Following this definition the constraint set of an edge can be written as $(C_e|D_e)$

R-Vines structures are not unique, the number of potential R-Vines rises extremely fast with the number of variables in the data set. Morales-Nápoles show that the number of possible R-Vines structures for n variables is $n! * 2^{\frac{n^2-5n+4}{2}}$.[15]

From the Proximity condition in definition 13 follows that every possible pair of variables occurs exactly once as the conditioned set of an edge in an

R-Vine. Because of this the R-Vine can be used as structure for the pair-copula-construction.
With the notation of the last definition the pair-copula-construction of the density for n random variables $f_{1,2,...n}$ connected by an R-Vine structure can be written as

$$f_{12...n} = f_1 f_2 ... f_n \prod_{i=1}^{n-1} \prod_{e \in E_i} c_{(j(e),k(e)|D(e))}(F_{(j(e)|D(e))}, F_{(k(e)|D(e))})$$

In this formula E_i denotes the set of edges in tree T_i we leave out the arguments for the density and distribution functions in the PCC , as they are defined by the same subscript as their respective function.

In the following sections we will give examples for two special cases of R-Vines and a small example for an R-Vine in order to introduce some notation, which is helpful to effectively handle R-Vine structures.

C-Vines and D-Vines

Two special cases for R-Vines are the C-Vine and the D-vine.

Definition 15. ***C-Vines and D-Vines***

- *A R-Vine with exactly two leaves in the first tree T_1 is called a D-Vine.*
- *A R-Vine for which every tree $T_i \quad i = 1, ...n$ contains exactly one node of degree n-i is called a C-Vine.*

The main feature of an D-Vine is that once the initial tree is selected the shape of the remaining trees is already determined by the proximity condition.
In an C-Vine every level will add the same variable to the conditioning sets of all edges in the next tree. As all edges in a tree share the central node in each tree of the C-Vine, all remaining nodes are always available for the central node when selecting the next tree.

In figure 2.6 we give a graphical example for both an C-Vine and an D-Vine, both for five random variables. The arrows in 2.6 represent the way edges of one tree turn to nodes of the next tree. We choose not to include all arrows for the C-Vine, as we believe that this would make the image less clear.

The C-Vine in our example also makes it clear that while the proximity condition is necessary for an R-Vine it is not mandatory that two edges sharing a node in one tree are directly connected by an edge in the consecutive tree.

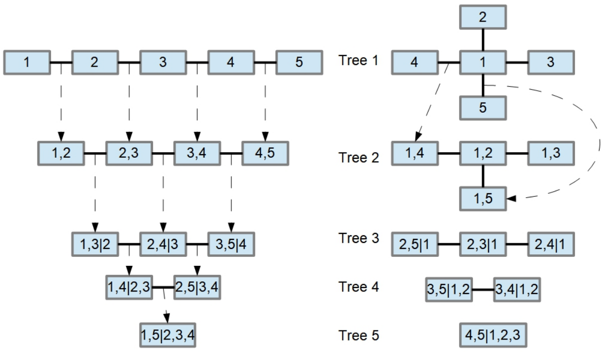

Figure 2.6: Both an D-Vine (left) and C-Vine (right) for five variables.

Example for an R-Vine

Up to this point we always presented an R-Vine structure by a graphical representation of it´s trees. This can become both quite demanding in space and incomprehensible to a reader as the number of variables and with it the size and complexity of the trees increases.
As the trees in the R-Vine are fully specified by the constraint sets of their edges a more economic method to give the structure of an Vine is the definition of an R-Vine array by J.F. Dissmann [14]. With this method the whole structure of the trees can be presented as a triangular array.

To give an example the R-Vine given by figure 2.7 can be represented by the array

$$\begin{pmatrix} 6 & & & & & \\ 1 & 3 & & & & \\ 4 & 1 & 1 & & & \\ \underline{5} & 4 & \underline{4} & 2 & & \\ 2 & 5 & 5 & 4 & 5 & \\ 3 & 2 & 2 & 5 & 4 & 4 \end{pmatrix}$$

.

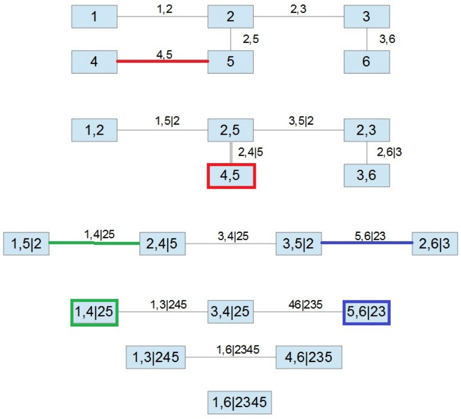

Figure 2.7: An R-Vine with 6 nodes

In this array for every tree T_i the conditioned set of it's edges can be found by looking at the diagonal elements and the $(n+1-i)th$ element in the same column. The conditioning sets associated to the edges with these conditioned sets is given by the elements in the respective columns below the $(n+1-i)th$ row. We highlight some examples for this with the three coloured edges and nodes in figure 2.7 and the correspondingly coloured elements in the array above.
It should be noted that this representation is not unique, for every R-Vine structure with n initial nodes there are 2^n different arrays describing the same tree structure of the R-Vine. Another example for the R-Vine in image 2.7 can be found in the appendix under A.3.

In the same fashion the copulas fitted to the edges and their parameters con be represented as diagonal arrays of size n. This way the whole R-Vine is fully specified by three arrays, the R-Vine array, giving the structure for the trees in the vine, the copula array, assigning each edge of the trees a copula, and the parameter array, containing the parameters of the used copulas.

For the R-Vine in our example the joint density function would be

$$\begin{aligned} f_{123456} =& f_1 * f_2 * f_3 * f_4 * f_5 * f_6 * c_{1,2}(F_1, F_2) * c_{2,5}(F_2, F_5) * c_{2,3}(F_2, F_3) * c_{3,6}(F_3, F_6) * \\ & c_{4,5}(F_4, F_5) * c_{1,5|2}(F_{1|2}, F_{5|2}) * c_{2,4|5}(F_{2|5}, F_{4|5}) * c_{2,6|3}(F_{2|3}, F_{4|3}) * c_{3,5|2}(F_{3|2}, F_{5|2}) * \\ & c_{1,4|25}(F_{1|25}, F_{4|25}) * c_{3,4|25}(F_{3|25}, F_{4|25}) * c_{5,6|23}(F_{5|23}, F_{6|23}) * c_{1,3|245}(F_{1|245}, F_{3|245}) * \\ & c_{4,6|235}(F_{4|235}, F_{6|235}) * c_{1,6|2345}(F_{1|2345}, F_{6|2345}). \end{aligned}$$

2.5 Truncation

The example in the last section demonstrates that even for a low number of variables the joint density can already take quite a complicated form and require the estimation of many parameters. As it is unavoidable that estimation errors are made when working with real data and the copulas in the later trees can only be estimated conditionally on the copulas in the first trees any early estimation error affects the fitting of the future copulas.

For this reason it is favourable to have the strongest dependence in the first few trees. If most correlation is covered by the first few trees, it is often possible to ignore any remaining small correlations in the later trees and allocate the Product copula to the edges in these higher indexed trees instead of attempting to fit another copula. This procedure is called truncating the R-Vine. A truncation significantly simplifies the distribution of the R-Vine copula and greatly reduces the number of parameters needing to be estimated.

There are several ways to decide which level of truncation is optimal given a data set. In the following we will discuss three options, two methods based on the Vuong-test and one option based on Bayesian statistics using directed, acyclic graphs.

Vuong-test

The Vuong-test was first introduced by Vuong [19]. In this work we will use a special case given by Brechmann [1].

Definition 16. *Vuong-test*
Let RV_1 and RV_2 be two different R-Vine models given by Θ_1, Θ_2, where Θ contains the tree structure, the fitted copulas and their parameters. Let the density functions of RV_1 and RV_2 be rv_1 and rv_2. Furthermore let the dataset consist of vectors of observations $o_i = (v_{1,i}, v_{2,i}, ...v_{n,i})$ of the random variables.
Define $d_i := ln(rv_1(o_i|\Theta_1)) - ln(rv_2(o_i|\Theta_2))$ as the difference in the log-likelihood of the observation given Θ_1, Θ_2.

Under the null hypothesis that both models are equally close to the true distribution the expected value of d is zero and the test statistic

$$v := \frac{\sum_{i=1}^{n} d_i}{n\sqrt{\sum_{i=1}^{n}(d_i - \bar{d})^2}}$$

should follow a standard normal distribution.

The Vuong-test compares two not necessarily nested models and determines if both are equally close to the true, unknown distribution of the data using the log-likelihood of the two models.

An R-Vine usually has many parameters that need to be estimated, in order to prevent over-fitting it is recommended to use some form of punishment term for the number of parameters in the model on the log-likelihood, analogous to the punishing terms for the Akaike(AIC)- or Schwarz(BIC)-criterion in model selection.

Even though the Vuong-test does not require the tested models to be nested, in this work we only apply the test to nested models. In order to determine appropriate truncation levels we use the Vuong-test in two different ways.

For the first method we first build an R-Vine truncated at level one. For this we simply set the copulas for the edges in all but the first tree to the Product copula. Because we consider all pairs of variables not occurring in the first tree independent the structure of the following trees is not important and we can halt the algorithm for determining the R-Vine. This R-Vine is tested with the Vuong-test against an R-Vine truncated at level two.
If the larger R-Vine is significantly better we dismiss the smaller R-Vine and build an R-Vine truncated at the next higher level.
We continue until the larger R-Vine is no longer a better fit. Once this happens we say that the truncation level of the smaller R-Vine is the optimal truncation level.

This method is computationally relatively efficient as we do not need to estimate the full R-Vine and use the Vuong-test for relatively small R-Vines. A drawback on the other hand is that since the proximity condition limits the possible pairs of variables considered in each tree it might not be always possible to use the available information in the dataset in an optimal way because variable pairs with high correlation might not be available to be immediately used in the current tree and can only be used in later trees. In such cases this method can terminate early so that the R-Vine is truncated at a too low level and significant information can be lost.

The second method to determine an optimal truncation level based on the Vuong-test compares the truncated models to a fully determined R-Vine.

First the full R-Vine is estimated from the data. Then we set the copula associated to the edge in the last tree to the Product copula and test this, truncated at level n-1, model against the full model. If it is not significantly worse we set the copulas in the second to last tree also to the Product copula and again test against the full, not truncated R-Vine.
We continue until the we arrive at a model that is significantly worse compared to the full model, and say that the optimal truncation level is the last truncation level that was not worse than the full model.

The benefit of this method is that it will not result in a too low truncation level and waste information. On the other hand for this method not only the full R-Vine must be fitted first which can become quite computationally expensive for larger datasets, but also the Vuong-test will be more involved as the tested models are far more complex compared to the first method and this increases the required processing power.

As both models do not scale well with the number of parameters in the dataset and the Vuong test requires relatively many observations to be conclusive, in the next chapter we discuss a third way to determine an appropriate truncation level that is not based on the Vuong-test.

Directed Acyclic Graphs

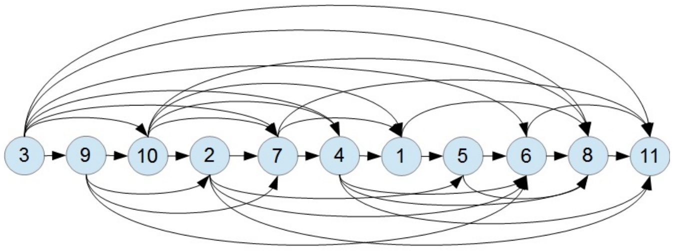

Figure 2.8: A directed, acyclic graph estimated from a subset of river-basins

This approach for finding an appropriate truncation level for the R-Vine structure utilises Bayesian networks. A Bayesian network represents the conditional dependencies in the given data in the form of a directed, acyclic graph in such a way that the *Markov property* holds. The *Markov property* describes that a node only depends on it´s parents and is conditionally on the parents independent of all it´s other ancestors.

C. Czado discusses the similarities of directed, acyclic graphs and truncated R-Vines .[4] Following the argumentation in this paper we try a method to establish an appropriate truncation level for a R-Vine that does not require the estimation of the R-Vine first.

Firstly we use an Tabu greedy search algorithm provided in the R-package bnlearn to try to find a Bayesian network representing the dependency structure for our data. We use the Tabu algorithm without any restrictions regarding the maximal number of parents. Then we search in the resulting directed, acyclic graph for the node with the highest number of parents and use this number as the truncation level for our R-Vine. The figure 2.8 gives an example for a dataset consisting of eleven variables. In it the node for variable "8" has the highest amount of parents, ("1","3","4","5","6","10").

2.6 Algorithms to estimate R-Vines

Copula selection

There are several possible ways to fit a copula to a pair of random variables.

The first method we consider is the maximum-likelihood method. For this method for every copula family that is considered the parameters maximising the likelihood of observing the relationship in the given data is calculated, and the copula with the highest maximum likelihood is selected to represent the dependency between the variables.

A second possible method to estimate copulas with one parameter utilises Kendall´s τ and tries to fit and select the copulas according to the empirical correlation in the given random variables.

Both methods can be applied for selecting among the copulas we consider in this work, however if one expands the range of considered copulas to copulas exhibiting different tail-dependence parameters $0 \neq \lambda_u \neq \lambda_l \neq 0$, which require at least two parameters, the method based on Kendall´s τ is no longer viable.

For this reason we will use the maximum-likelihood method for the remainder of the work to select the copulas for our R-Vines.

In addition we perform a test for independence before we attempt to fit a copula to an edge of a tree. We do this both to simplify our R-Vine-copulas and to avoid estimation errors as the selection from our copula families is especially hard when the variables are close to independent because members of all our families can be arbitrarily close to the product copula.

In the following we present three different approaches for the estimation of R-Vine models. The first two methods that will be showcased work well together with our presented methods of truncation and our data, the third method by Kurowicka works somewhat different and is better suited for other applications.

Goodness of fit-method

This method is described in a paper by C. Czado. [9]
This method aims at generating the best fitting trees in each step in order to minimize estimation errors when deciding about the later trees and decides which edges should belong to the individual trees accordingly.

1. Start with the complete graph with $\binom{n}{2}$ edges,
2. Fit a copula $C_{i,j}$ to every edge (v_i, v_j) of the graph,
3. Calculate the Akaike Information Criterion AIC for all copulas and use -AIC as the weight for the corresponding edges,
4. Apply a Maximum-Spanning-Tree algorithm to the graph,
5. Use the edges of the resulting tree as nodes for a new complete graph,
6. Delete all edges in the new graph which do not fulfil the Proximity condition from definition 13,
7. Use the fitted copulas to transform the variables v_i into $v_{i|j}$ to account for correlation explained by the copulas $C_{i,j}$,
8. Continue from step 2 onward until the full R-Vine is specified.

The Akaike Information Criterion in the above algorithm can be replaced by other goodness-of-fit measures like the log-likelihood.

This method cares less about the correlation in the data-sample compared to the following algorithms, it is well suited in situations when there is little information about possible causes for correlation in the data available. As this algorithm fits the copulas before it decides on the tree structure it can require a lot of computational power to calculate copulas that are dismissed in the next step of the algorithm.

Sequential method

This algorithm was introduced by Dissmann. [14] The sequential method cares less about generating the best fitting tree in each step but rather tries to find the R-Vines which capture the highest correlation between variables in the early trees. It is recommended to use Kendall's τ as a measure for the correlation between the random variables.

- Start with the complete graph with $\binom{n}{2}$ edges,
- Calculate Kendall's τ for each edge and use the absolute value of it as the weight for the corresponding edge,
- Apply a Maximum-Spanning-Tree algorithm to the graph,
- Fit a copula $C_{i,j}$ to every edge (v_i, v_j) of the resulting tree,
- Use the edges of the resulting tree as nodes for a new complete graph,
- Delete all edges in the new graph which do not fulfil the the Proximity condition from definition 13,
- Use the fitted copulas to transform the variables v_i into $v_{i|j}$ to account for correlation already explained by the copulas $C_{i,j}$,
- Continue from step 2 onward until the full R-Vine is specified.

The algorithm can be adjusted to use a different weight function, like different measures of correlation or predetermined connections (in regard to our topic this can be natural river connections) when selecting the tree.

This algorithm is very similar to the first algorithm we presented, the main difference is that the copulas are fitted before the selection of the trees. However this algorithm is more efficient compared to the last presented algorithm as it avoids to calculate unnecessary copulas in each step and only fits those that will remain in the model.
Together with the flexibility in the weight function this makes this algorithm

the standard algorithm when fitting R-Vines and we will use it for the remainder of this work.

The first two algorithm we presented are the same in the respect that one starts with the first tree of the R-Vine and works forwards until the full R-Vine is specified and both decide both the shape of the trees and the copulas for the edges of the tree in the same round of the algorithm. In the next section we will shortly present an algorithm that differs from them in both this regards.

Kurowicka‘s algorithm

This algorithm was proposed by Kurowicka [13]. Unlike the previous two algorithms presented this algorithm starts with the solitary node of the last tree T_n and works it's way back to the first tree T_1 while always choosing edges with the least possible partial correlation so that the first trees.
In order to ensure that the result will be an R-Vine, Kurowicka first introduces the following notation.

For an edge in the form $e = (x, y|A_e)$ with conditioned set $\{x, y\}$ and conditioning set A_e we define $B_x := \{x\} \cup A_e$ and $B_y := \{y\} \cup A_e$.

Using this notation, the following conditions ensure that the at the end of the algorithm an R-vine will be specified.

- **Condition 1** If in tree T_j there are an x and y such that $B_x = B_y$ and $x \neq y$ then there must me a node in tree T_{j-1} with the conditioned set $\{x, y\}$.

- **Condition 2** For all $B_{i_1}, ...B_{i_k}$ such that $|B_{i_p} \triangle B_{i_q}| = 2$

  $$B_{i_p} = \{i_p, s|A_{i_p} \setminus \{s\}\} \vee B_{i_p} = \{i_p, t|A_{i_p} \setminus \{t\}\}, \quad s, t \in A_{i_p}$$

  $\triangle$ denotes the symmetric difference, the union without the intersection, of two sets. $A \triangle B := (A \cup B) \setminus (A \cap B)$.

With these two conditions the Algorithm by Kurowicka works the following way.

1. Choose two variables $x, y \in 1, 2, ...n = I$ as conditioned set for the sole edge in T_{n-1}, choose partners $p(x), p(y)$ for x, y in $I \setminus \{x, y\}$ to get the two edges $(x, p(x)|I \setminus \{x, y, p(x)\})$ and $(y, p(y)|I \setminus \{x, y, p(y)\})$ of T_{n-2},

2. Find the sets B_i for all edges of the last tree T_j,

3. Remove sets for which $i_p = i_q$ for $p \neq q$,
4. Apply Condition 1,
5. Choose partners for the variables satisfying Condition 2 for the tree T_{j-1} ,
6. Go back to step 2 until the R-Vine tree structure is fully specified.
7. Fit copulas to the trees.

This algorithm provides an alternative to the previous forward working algorithms. As argument for the choice amongst available partners in step 5, Kurowicka suggest the partial correlation between the variables.

This makes this algorithm especially suited when working with elliptical copulas, but unsuited for our selection of copulas. It is also not an ideal algorithm when one wants to study the dependence for extreme events as the only elliptical copula that we consider for this work does not exhibit tail-dependence as can be seen in table A.2.
We can observe again that the restrictions of the R-Vine structure might prevent the algorithm from finding the R-Vine-structure that is optimal for truncation.

Chapter 3

Data

In this section we will discuss the data that we use to estimate the dependency structure, as well as the data we use for our forecasts.

The main source of our data is the Joint Research Centre of the European Union. Our data consist of two separate parts, historic data for the peak water discharges in European rivers and estimations of future economic losses in loss-areas associated to these rivers.
In the following two sections both types of data will be described.

3.1 Historic Data for the water discharge

We have monthly data for both the maximal and the median water discharge for 776 riverbeds across Europe available for the estimation of our models. The observation period for these water discharges ranges from the year 1990 until the year 2011.
In addition to the data on the water discharge in the river beds we also have the geographic locations of the rivers and data on natural connections between the rivers available.

The figure 3.1 gives a graphical representation of the the river basins we use as the basis for this work.
Because in this work we want to analyse dependency under extreme events, it seems reasonable to us to use the data for the peak instead of the median water discharge to estimate our R-Vines.

As we have relatively few observations for each river basin available we decide for this work to forgo controlling for seasonal effects, as this would require us to sacrifice additional twelve observations for each random variable.

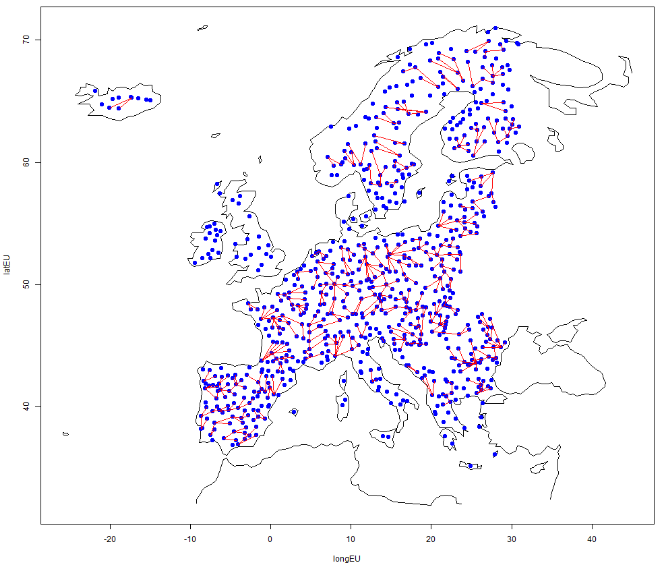

Figure 3.1: River-basins are represented by the blue dots, a connection between two rivers is given by a red line.

However we want to note that we verified that controlling for seasonality does not greatly impact the results presented in this work and could be easily implemented without invalidating any of our presented results if enough data becomes available.

Our first step is to verify whether there is significant correlation in the river basins´ water discharges.
For this we calculate a correlation coefficient, throughout this work we will use Kendall´s τ as standard, for each of the possible $\binom{776}{2} = 300700$ pairs of variables.
As can be seen in the image 3.2 even though most pairs are little correlated or even uncorrelated, there are many highly correlated river beds in the data.

This suggests that we can separate our variables so that rivers in the resulting smaller sets exhibit above average correlation with other members of their own set, while rivers belonging to different sets are uncorrelated from each other.

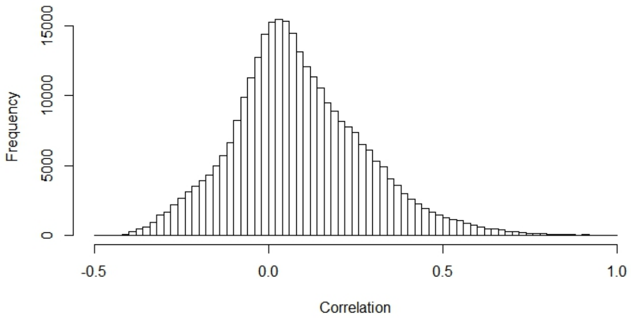

Figure 3.2: A histogram of the correlation found in the JRC data

Clustering

In order to find these smaller, highly correlated subgroups in the data set, we apply a hierarchical clustering algorithm.
For a hierarchical clustering initially each random variable is considered a separate cluster. The algorithm then merges the two clusters with the smallest distance between them in each step until the desired number of clusters is reached.
We tried several different metrics for the initial distance function and also various distinct methods to update the distance between newly merged and old clusters and found that a combination of the information distance, as we defined it in chapter 2, and updating our distances analogous to Ward´s method during the algorithm gives the most satisfying results.

Definition 17. ***Ward's method***
Let C_i, C_j, C_k be three clusters with cluster-sizes n_i, n_j, n_k and let the distance between two clusters C_i and C_j be denoted by the function $d_{i,j}$. Define $C_l := C_i \cup C_j$. Then Ward´s method written in the Lance-Williams recursive formula [21] is

$$d(C_l, C_k) = \frac{n_i + n_k}{n_l + n_k} d(C_i, C_k) + \frac{n_j + n_k}{n_l + n_k} d(C_j, C_k) - \frac{n_k}{n_l + n_k} d(C_i, C_j).$$

In the original paper [11] Ward merges clusters such that the sum of the variances of the clusters is minimal. This corresponds to the recursive formula with the initial distance defined as $d(X, Y) = ||X - Y||^2$. Because of this it is also known as Ward´s minimum variance method.

For deciding on the optimal number of clusters we looked at several different possible criteria like the Dunn index [5] or the Calinski and Harabasz criterion [8]. We find that sixteen clusters provide a reasonable trade-off between separating our data in easier to manage subgroups and keeping most correlation between the individual river beds in the data.

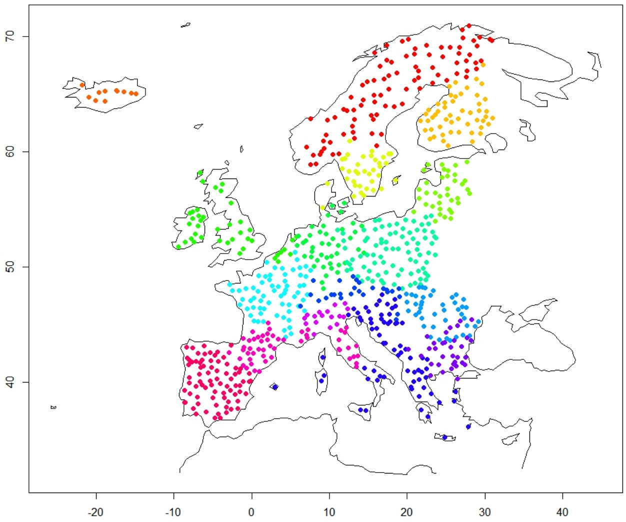

Figure 3.3: The clustering for Europe

The results of this clustering approach can be seen in figure 3.3 and table A.4. We feel it is important to note that the clustering algorithm only received data for the water discharge but no geographic information. The fact that all clusters are geographically concentrated and that there is little overlap between clusters confirms the assumption of climate simultaneously influencing the behaviour of river beds in larger areas.
The size of our individual clusters is given in table 3.1.

Cluster ID	1	2	3	4	5	6	7	8
Number of basins	89	11	51	37	37	41	43	85

Cluster ID	9	10	11	12	13	14	15	16
Number of basins	65	43	26	70	33	17	52	76

Table 3.1: An overview of the clusters

In the following we will treat these sixteen clusters as independent from each other to simplify further calculations. This will lead to a lower forecast for the total damages in Europe since there is still some positive correlation between river beds in different clusters left, this is especially true for rivers close to the borders between clusters.

3.2 Provided Forecasts

As forecast for future losses we use data provided by the LISFLOOD model of the European Union's Joint Research Centre (JRC). The LISFLOOD model is a hydrological rainfall and water run-off model, capable of simulating the hydrological processes that occur in a catchment .[17] With the help of the LISFLOOD model and various future scenarios the JRC estimated flood damages in 1433 different loss areas. These loss areas can each be connected to at least one of the rivers we use for the estimation of the R-Vines.
The estimation by the JRC are provided in the form of yearly return periods ranging from 1995 to 2085.

Definition 18. ***Return period***
A return period RP_t for an event of magnitude k is the expected time, usually in years, between two occurrences of the event. It is related to a quantile q_α from the distribution function of possible magnitudes within a year by

$$\alpha = 1 - \frac{1}{RP_t}.$$

The range of provided return periods and the corresponding quantiles can be seen in table 3.2.

RP	2	5	10	20	50	100	250
α	.5	.8	.9	.95	.98	.99	.996
RP	500	750	1000	2500	5000	7500	10000
α	.998	$.998\dot{6}$	.999	.9996	.9998	$.9998\dot{6}$	.9999

Table 3.2: Return Periods and corresponding quantiles

Actual losses larger than zero are provided for 914 of the 1433 areas. The lack of forecasts for this many loss areas causes that several of our 776 river basins cannot be linked to at least one loss-area with positive estimated losses. We nonetheless decide to keep these river basins in our R-Vine models, because we are of the opinion that even though their presence does not directly affect the total predicted losses in an cluster, they are still valuable in regards to the estimation of dependency structure between the river beds.

Additionally this way the predictions for the total losses can be easily updated if new forecasts for losses associated to these rivers are provided.

The forecast coming from the JRC is generated by first simulating the losses for some years and calculating key quantiles, or return periods, from these simulations. They use a thirty year timespan between the years in their simulation. The quantiles for the remaining years are subsequently estimated through interpolation.

This generates two different problems. The larger concern are cases of missing data, an example for an loss-area with two forecasts initially missing is loss-area 36 in figure 3.4. As can be seen in this image the missing values do not only generate a problem for the directly affected year but spread through the periods before and after.
A second source for errors is that the interpolating polynomials do not necessarily show the same monotonicity as the simulated points. An example for this can be seen by loss-area 283 in figure 3.4, the 0.99 quantile represented by the violet line intersects and for some years exceeds the 0.996 quantile represented by the dashed black line.

As both this types of errors are rare in the provided forecasts, only six loss areas are affected by one of them, we decide not to remove the affected loss areas and instead attempt to correct them. We tried to alter the data as little as possible in doing so. For the first type of error we replace the polynomial interpolation during the affected years by a linear combination of the neighbouring forecasts. In the cases when the interpolation of an quantile by the JRC gives a result which is contradicting the axioms of probability we try to estimate the wrong quantile by the next higher and next lower quantiles.

The corrections for the two examples we provided is given in image 3.5.

The JRC provides fourteen different quantiles from the distribution functions of damages in each loss-area for the years 1995-2085. For our work we need the full distribution of the damages. In order to find the quantile function we decide to interpolate the given quantiles with cubic splines. It is important to pay attention to the fact that the result of an unrestricted interpolation is not necessarily monotone, even if the initial points are strictly monotonic. To avoid non-monotone results we apply the method described by Fritsch and Carlson [2] to restrict our interpolating polynomials sufficiently to ensure the required monotonicity.
Figure 3.6 gives an example of a cumulative distribution function derived

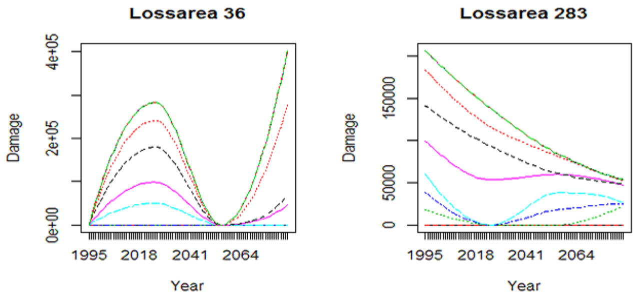

Figure 3.4: Two examples for errors in the forecast

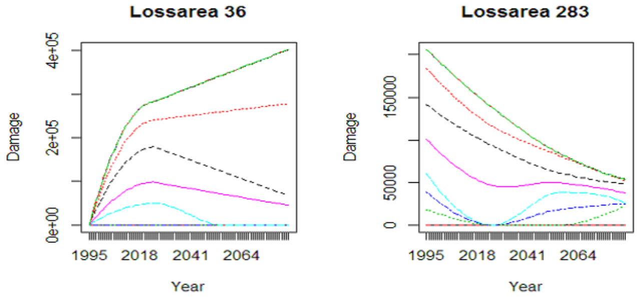

Figure 3.5: Our corrections for our two examples

this way, for both the whole range of damages and the last decile.

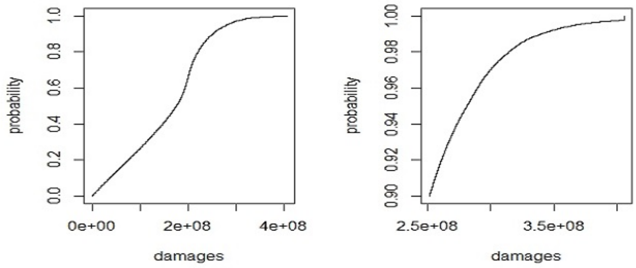

Figure 3.6: The cdf for the damages predicted for loss area 8 in the year 2050

In addition to the data from the JRC we also use protection standards for the individual loss areas provided by Jongman et al. [7]. The protection standards are given in the form of return periods and give the severity upon which damages caused by an event are negated through protective measures.

Chapter 4

Vines

Out of the three algorithms for the estimation of R-Vines we presented in chapter 2 we decided to use Dissmann´s sequential algorithm as our main interest lies in the dependency exhibited in the trees with the lower indices and Dissmann´s algorithm allows us to prioritize this while also being the most efficient algorithm for our data.
We use the absolute value of Kendall´s τ as weight for the edges for the selection of the trees in the algorithm.

The first tree for each cluster following Dissmann´s sequential algorithm can be seen in image 4.1. It is noticeable that most edges are short so that mostly pairs of geographical close rivers appear as edges in the first trees of the R-Vines, which is in line with initial assumptions.

In the next section we take a closer look at one example from our fitted R-Vines.

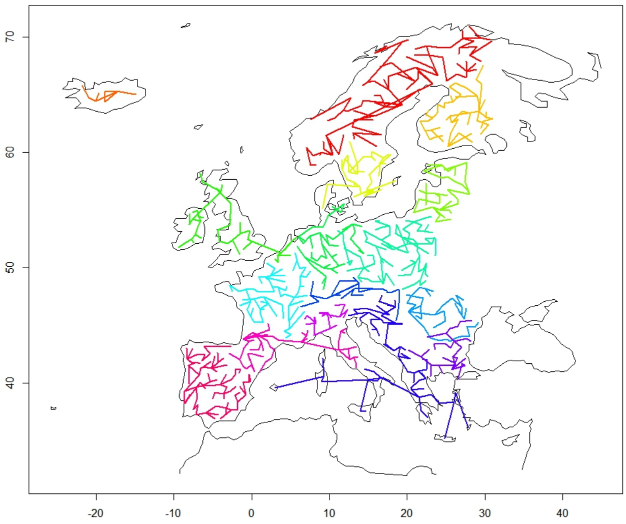

Figure 4.1: The first tree for each cluster

4.1 Cluster 2-Iceland

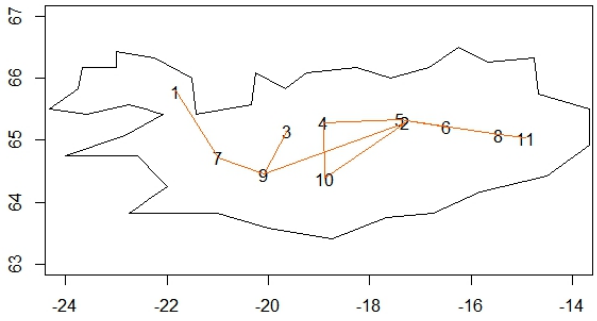

Figure 4.2: Iceland or Cluster 2 in our clustering

In order to give an example for the fitted R-Vines in this chapter we present an R-Vine for one cluster in greater detail. As most clusters are too big to serve as an easily understood example we decide to use our smallest cluster, which covers Iceland and is number two in our list of clusters, for this example.

A secondary reason to choose this cluster is that it is well disconnected from the other clusters. This can be easily verified by investigating the correlation of river beds in this cluster and the rivers in the remaining fifteen clusters. Figure 4.3 gives a histogram of this correlation. Most remaining correlations are in the range [-0.15,0.05] which allows to assume independence between the two sets of river basins .

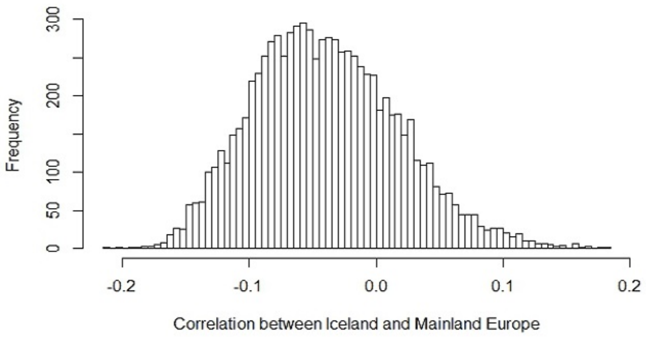

Figure 4.3: A histogram of the pairwise Kendall's τ for the rivers of Iceland and the rest of Europe. We calculated Kendall´s τ for all possible $8415 = 11 * 765$ combinations.

As this cluster two only contains eleven individual river beds, it is a ideal candidate to showcase the structure of an R-Vine and our three methods of truncation.

The first tree of this cluster is given in figure 4.2. In this image we also give our labelling of the individual rivers that we will use for the remainder of this example.

Using the notation we presented in chapter 2 the trees of this R-Vine are given by the array in table 4.1.

The copula families fitted to the edges of the trees given by 4.1 are given by the array 4.2 and the corresponding parameters for the copulas by the array 4.3.

$$\begin{pmatrix}
6 & & & & & & & & & & \\
3 & 11 & & & & & & & & & \\
1 & 3 & 8 & & & & & & & & \\
7 & 1 & 3 & 5 & & & & & & & \\
9 & 7 & 1 & 3 & 1 & & & & & & \\
2 & 9 & 7 & 1 & 3 & 3 & & & & & \\
10 & 2 & 9 & 7 & 4 & 4 & 7 & & & & \\
4 & 10 & 2 & 9 & 10 & 7 & 4 & 9 & & & \\
11 & 4 & 10 & 2 & 2 & 10 & 10 & 4 & 2 & & \\
8 & 5 & 4 & 10 & 9 & 2 & 2 & 10 & 4 & 4 & \\
5 & 8 & 5 & 4 & 7 & 9 & 9 & 2 & 10 & 10 & 10
\end{pmatrix}$$

Table 4.1: The R-Vine array for cluster 2

$$\begin{pmatrix}
. & . & . & . & . & . & . & . & . & . & . \\
\text{P} & . & . & . & . & . & . & . & . & . & . \\
\text{P} & \text{Ga} & . & . & . & . & . & . & . & . & . \\
\text{P} & \text{P} & C^{270} & . & . & . & . & . & . & . & . \\
Gu^{90} & \text{P} & \text{F} & \text{P} & . & . & . & . & . & . & . \\
\text{P} & \text{P} & \text{P} & C^{180} & \text{P} & . & . & . & . & . & . \\
\text{Ga} & Gu^{180} & Gu^{90} & Gu^{180} & \text{Ga} & \text{P} & . & . & . & . & . \\
\text{P} & \text{P} & \text{P} & \text{F} & \text{P} & \text{F} & \text{P} & . & . & . & . \\
C^{270} & C^{90} & \text{Gu} & \text{Gu} & \text{P} & \text{Gu} & \text{F} & \text{Ga} & . & . & . \\
\text{C} & \text{P} & Gu^{270} & Gu^{180} & \text{J} & \text{F} & Gu^{90} & Gu^{90} & \text{P} & . & . \\
\text{Gu} & \text{Gu} & \text{Gu} & \text{Gu} & \text{Gu} & \text{Gu} & \text{Gu} & \text{Gu} & \text{Gu} & \text{Gu} & .
\end{pmatrix}$$

Table 4.2: The fitted copulas for cluster 2, The letter P stand for the independence respectively the Product copula, Ga for a Normal or Gaussian copula, C for the Clayton copula, Gu for the Gumbel copula, F for the Frank copula and J for the Joe copula. A superscript, if present, denotes by how many degrees the copula is rotated

4.2 Truncation

The directed, acyclic graph for this cluster is already presented in chapter 2 as image 2.8.
The results of using the Tabu-algorithm to find the optimal truncation levels for the remaining clusters in our work are given by table 4.4.

$$\begin{pmatrix}
. & . & . & . & . & . & . & . & . & . & . \\
0.0000 & . & . & . & . & . & . & . & . & . & . \\
0.0000 & -0.24286 & . & . & . & . & . & . & . & . & . \\
0.0000 & 0.00000 & -0.21597 & . & . & . & . & . & . & . & . \\
-1.2386 & 0.00000 & 1.20045 & 0.00000 & . & . & . & . & . & . & . \\
0.0000 & 0.00000 & 0.00000 & 0.37722 & 0.00000 & . & . & . & . & . & . \\
0.2166 & 1.24362 & -1.17633 & 1.12945 & 0.45045 & 0.0000 & . & . & . & . & . \\
0.0000 & 0.00000 & 0.00000 & 1.24980 & 0.00000 & 1.8335 & 0.00000 & . & . & . & . \\
-0.2676 & -0.32260 & 1.10723 & 1.31903 & 0.00000 & 1.1837 & -4.89764 & 0.26564 & . & . & . \\
1.1320 & 0.00000 & -1.41912 & 1.30756 & 1.15489 & 2.4636 & -1.22729 & -1.19087 & 0.0000 & . & . \\
4.7861 & 8.71813 & 5.19527 & 6.55273 & 6.77228 & 10.1681 & 8.64628 & 7.74405 & 10.5221 & 6.8506 & .
\end{pmatrix}$$

Table 4.3: The parameters for fitted copulas for cluster 2

Cluster ID	1	2	3	4	5	6	7	8
Trunc. level	11	6	12	12	9	7	10	12

Cluster ID	9	10	11	12	13	14	15	16
Trunc. level	9	7	7	10	9	7	10	13

Table 4.4: The optimal truncation level for clusters with the DAG method

For cluster two the approach with directed acyclic graphs suggests a truncation level of six.
Alternatively the two methods based on the Vuong-test in chapter 2 yield the results in the tables 4.5 and 4.6.

In table 4.5 we test truncated Vines against the full model. In addition we choose to use a correction based on the Schwarz information criterion to account for the number of parameters in the R-Vines. The p-value in this table gives the probability that both tested model represent the data equally well. Depending on which p-value is deemed acceptable, the optimal truncation level with this method ranges from seven to eight which is higher or equal to the truncation level attained via the directed, acyclic graph method. Therefore this truncation method suggest more complex models with a better fit compared to the method utilising directed, acyclic graphs.

In the table 4.6 we use the Vuong test in the opposite direction, we start with the R-Vine model in which only the first tree contains copulas other than the Product copula and always test a model against the model with

trunc. levels	10vs8	10vs7	10vs6	10vs5	10vs4	10vs3	10vs2	10vs.1
test-stat.	2.004	2.515	3.423	4.223	5.800	6.169	7.767	8.713
p-value	0.045	0.012	0.001	2.4e-5	6.6e-9	6.8e-10	8.0e-15	0
Schwarz stat.	1.396	1.704	2.512	3.233	4.595	4.761	6.287	7.262
adj. p-value	0.163	0.088	0.012	0.001	4.3e-6	1.9e-6	3.2e-10	3.8e-13

Table 4.5: The Vuong-test testing various truncation levels,given by the number of trees included in the model, against the full model with the all ten trees

one additional tree containing copulas other than the Product copula. With this method the first suggested truncation level is three, well below the truncation level suggested by the other two methods.

trunc. levels	1vs2	2vs3	3vs4	4vs5	5vs6	6vs7	7vs8	8vs9
test-stat.	-5.352	-6.304	-1.994	-4.564	-2.575	-2.692	-1.878	-2.004
p-value	8.7e-8	2.9e-10	0.046	5.0e-6	0.010	0.007	0.060	0.045
Schwarz stat.	-4.689	-5.368	-1.097	-3.712	-2.166	-2.082	-1.231	-1.395
adj. p-value	2.8e-6	8.0e-8	0.273	2.1e-4	0.030	0.037	0.218	0.163

Table 4.6: The Vuong-test testing R-Vine testing mo.

In both tables we do not test a model with 10 trees against a model with 9 trees because the copula assigned to the final edge is the Product copula, as can be seen in array 4.2, so that truncation at this level does not change the model.

The methods based on the Vuong-test are computationally expensive for our larger clusters. Additionally we only have 264 observations, a number that for the larger clusters is dwarfed by the number of parameters to be estimated, our largest cluster containing 89 rivers requires the selection 3916 copulas and their parameters. Considering that this affects the power of the Vuong-test for the rest of this work we use the truncation levels derived by utilizing directed, acyclic graphs which are given in table 4.4.

Chapter 5

Forecast

In this chapter we first present the algorithm used to calculate the joint distribution function for the sum of total damages in a cluster and then introduce two risk measures the Value-at-Risk and the Expected Shortfall. Afterwards we take a closer look at the loss forecast for one of the clusters. In the last part of the chapter we will provide the Value-at-Risk at different levels α for our clusters in two sample years, the year 2020 and 2085, as well as the expected shortfall for the same levels in these years.

Since our marginal distributions are only piecewise parametric, which greatly increases the complexity of the joint density function of our random variables and we are more interested in the sum of total damages we decided to use the following algorithm to approximate the cumulative distribution function numerically instead of trying to analytically solve the joint density function.

- Generate a grid of possible losses,
- Generate m*n, where n is the size of the cluster, independent samples from a uniform distribution on [0,1].
- Use the R-Vine to transform this into m samples of a vector $\mathbf{v} = (v_1, ..., v_n)$ containing the severity of the losses associated to each river basin.
- Calculate the losses associated to each basin with the help of the quantile functions for the loss areas connected to the river basins $\mathbf{d} = (d_{1,1}(v_1), d_{1,2}(v_1), ..., d_{1,k}(v_1), d_{2,1}(v_2), ..., d_{n,k_n}(v_n))$.
- If protection standards are used set damages below the severity threshold to zero.

- Calculate the total damages $d_{1,1}(v_1) + d_{1,2}(v_1) + ... + d_{n,k_n}(v_n)$.
- Build the empirical cumulative distribution function.
- Evaluate the empirical cumulative distribution function on the grid-points.

Limitations to computational power, especially for larger R-Vines, restrict the number of samples m that can be drawn in one step. To circumvent this we decided to use the above algorithm several times on the same grid and then calculate the arithmetic mean of the individual functions. To give the numbers we choose m equal to 2000 and applied the algorithm 2000 times, resulting in 4000000 samples for each cluster in a year. We decided to use a grid with 100000 grid-points, with 10000 points in both the lowest and highest 5 percent of potential damages each.

In order to evaluate the risk arising from the dependency between the water discharge of our river beds we need risk measures. Since we are mostly interested in the behaviour at the right tail of the distribution we use the following definition for two possible risk measures to judge the risk presented by a distribution.

Definition 19. ***Value-at-Risk, Expected Shortfall***

The value at risk for the threshold α is defined as

$$VaR_\alpha := inf\{l \in \mathbb{R}, P(L > l) \leq 1 - \alpha\}$$

The expected shortfall, or average value at risk, is defined as

$$ES_\alpha := E[L|L > VaR_\alpha] := \frac{1}{1-\alpha}\int_\alpha^1 VaR_\gamma d\gamma$$

With this definition the VaR_α is another name for the α quantile, while the expected shortfall is the the expected value of the distribution conditional that a certain threshold severity is exceeded. The Value at Risk is the more intuitive of our two used risk measures, but unlike the Expected Shortfall it does not contain information how the distribution behaves past α. Therefore we consider the Expected Shortfall the superior risk measure as it gives the expected value of the magnitude of the extreme events and can therefore serve as an guideline for appropriate reserves.

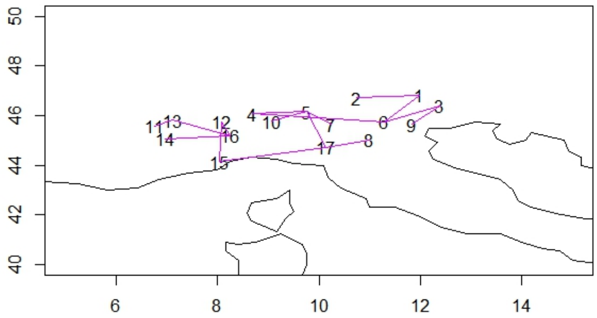

Figure 5.1: Cluster 14

5.1 Cluster 14-Northern Italy

Since the JRC data lacks any forecasts for Iceland, for this section we use our second smallest cluster depicted in figure 5.1. This cluster covers the northern part of Italy and the maximal loss according to the JRC forecast for this cluster in the year 2020 is $3.49 * 10^{10}$ Euro and $4.09 * 10^{10}$ Euro in the year 2085.

We first give a comparison between our R-Vine model C_{Vine} and assuming the water discharge of the rivers in the cluster as independent without the use of protection levels. Figure 5.2 gives a visual comparison of the cumulative distribution functions for the two models. Both models yield the same expected value however compared to the model with independence C_{Vine} suggests that both very high and very low damages occur more frequently.

The table 5.1 gives the Value-at-Risk and the expected shortfall for some sample levels. For the right tail of the distribution the Value-at-Risk for the C_{Vine} is 54.85 percent higher compared to independence, while the Expected Shortfall is 55.35 percent higher.

From now on we always use the provided protection levels. If the protection standards are used the cumulative distributions change to the functions shown in image 5.3 and table 5.2.

From the images 5.2, 5.3 and the tables 5.1 and 5.2 it becomes clear that if the simplifying assumption of independence is made even though the true distribution is close to our C_{Vine}, several errors are occur.

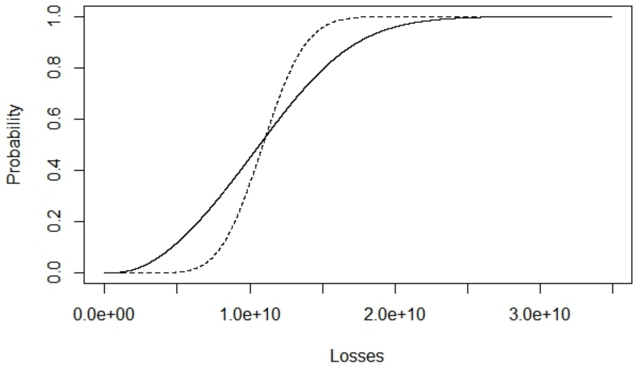

Figure 5.2: The predicted losses for the year 2020 in cluster 14 according to the R-Vine (solid) and assuming independence (dashed)

α	.9	.95	.975	.99	.999
VaR					
Indep.	$1.39 * 10^{10}$	$1.48 * 10^{10}$	$1.56 * 10^{10}$	$1.65 * 10^{10}$	$1.84 * 10^{10}$
C_{Vine}	$1.75 * 10^{10}$	$1.94 * 10^{10}$	$2.12 * 10^{10}$	$2.33 * 10^{10}$	$2.76 * 10^{10}$
ES					
Indep.	$1.51 * 10^{10}$	$1.59 * 10^{10}$	$1.65 * 10^{10}$	$1.74 * 10^{10}$	$1.92 * 10^{10}$
C_{Vine}	$2.01 * 10^{10}$	$2.18 * 10^{10}$	$2.34 * 10^{10}$	$2.53 * 10^{10}$	$2.88 * 10^{10}$

Table 5.1: Value at risk and Expected Shortfall for levels α for both models.

α	.9	.925	.95	.975	.99	.999
VaR						
Indep.	$2.46 * 10^7$	$4.81 * 10^8$	$1.39 * 10^9$	$2.97 * 10^9$	$5.57 * 10^9$	$6.42 * 10^9$
C_{Vine}	$1.92 * 10^6$	$2.11 * 10^7$	$1.63 * 10^8$	$3.06 * 10^9$	$6.50 * 10^9$	$1.62 * 10^{10}$
ES						
Indep.	$2.10 * 10^9$	$2.71 * 10^9$	$3.61 * 10^9$	$4.89 * 10^9$	$6.16 * 10^9$	$8.04 * 10^9$
C_{Vine}	$2.11 * 10^9$	$2.81 * 10^9$	$4.20 * 10^9$	$7.08 * 10^9$	$1.03 * 10^{10}$	$1.88 * 10^{10}$

Table 5.2: Value at risk and Expected Shortfall for some α for both models.

The first error is that independence underestimates the probability of low impact events and the probability of high events, so that if real data is used to estimate the function the fitted distribution will be shifted one side and therefore have a too low or too high expected value, affecting all estimations

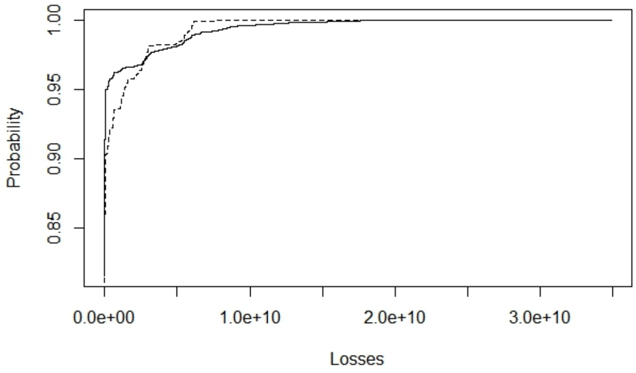

Figure 5.3: The cumulative distribution functions for the R-Vine copula (solid) and the independence model (dashed) for the losses in cluster 14 for the year 2020, if the protection standards are used

based on this distribution. Even if the first error can be avoided and the true expected value is reached, the second error is that the distribution under the assumption of independence has too thin tails in comparison to the true distribution which even with the true or a too high expected value leads to an underestimation of the risk when independence is assumed.

Both this effects are responsible that, if the dependence within the river system is ignored, there will not be sufficient reserves for the coverage of flood damages.

In the next step we look into how the distribution of damages changes over the years. In order to to this we use our algorithm to calculate the distribution functions for three years, 2020, 2055 and 2085. Figure 5.4 and table 5.3 give the corresponding visualisation and data. With our estimations we find that the expected shortfall, which we consider a good guideline for the amount of required risk reserves, increases by more than 3000 million Euro alone in the region covered by cluster 14 in the next seventy to sixty years.

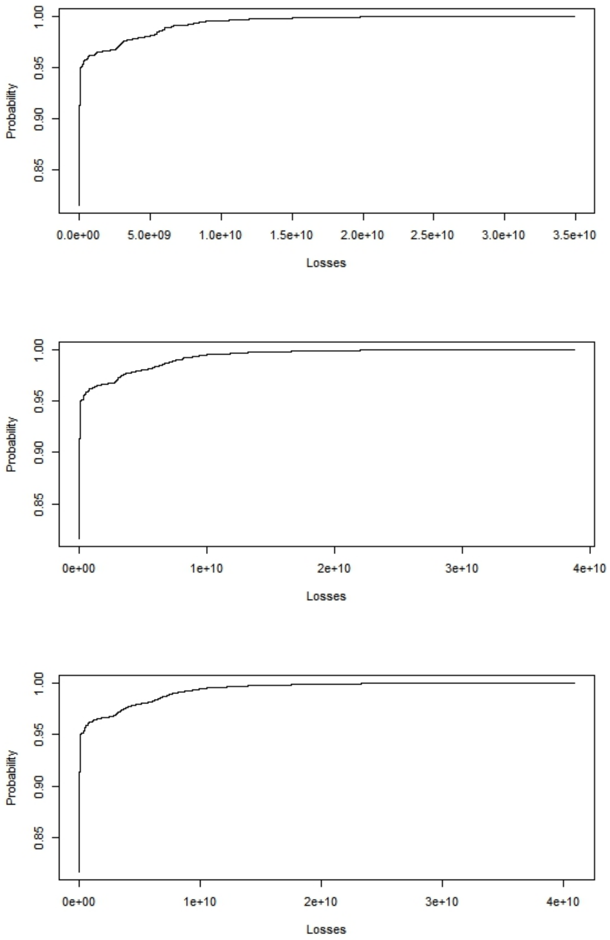

Figure 5.4: Cluster 14, for the years 2020 (top),2055(middle) and 2085 (bottom)

α	.9	.925	.95	.975	.99	.999
Value at Risk						
2020	$1.92 * 10^6$	$2.11 * 10^7$	$1.63 * 10^8$	$3.06 * 10^9$	$6.50 * 10^9$	$1.62 * 10^{10}$
2055	$1.94 * 10^6$	$2.15 * 10^7$	$1.69 * 10^8$	$3.33 * 10^9$	$7.49 * 10^9$	$1.76 * 10^{10}$
2085	$2.25 * 10^6$	$2.27 * 10^7$	$1.78 * 10^8$	$3.64 * 10^9$	$7.73 * 10^9$	$1.86 * 10^{10}$
Expected shortfall						
2020	$2.11 * 10^9$	$2.81 * 10^9$	$4.20 * 10^9$	$7.08 * 10^9$	$1.03 * 10^{10}$	$1.88 * 10^{10}$
2055	$2.33 * 10^9$	$3.10 * 10^9$	$4.64 * 10^9$	$7.86 * 10^9$	$1.14 * 10^{10}$	$2.07 * 10^{10}$
2085	$2.45 * 10^9$	$3.26 * 10^9$	$4.88 * 10^9$	$8.25 * 10^9$	$1.20 * 10^{10}$	$2.19 * 10^{10}$

Table 5.3: Some risk measures for cluster 14 in the years 2020, 2055 and 2085

5.2 Forecast for all clusters

In tables 5.4 and 5.5 it is possible see that climate change will increase the potential flood damages in some areas while in other clusters the losses due to flooding are decreasing in the future. The clusters with decreasing damages are located in north-east Europe, while damages in clusters in the south-west of the continent increase as can be seen in Figure 5.5. This can be explained because river floods in the north eastern regions are mainly caused by the spring snow melt which is predicted to decline, while the south-western region is more vulnerable to heavy rainfall which should increase due to the increased evaporation of ocean water.

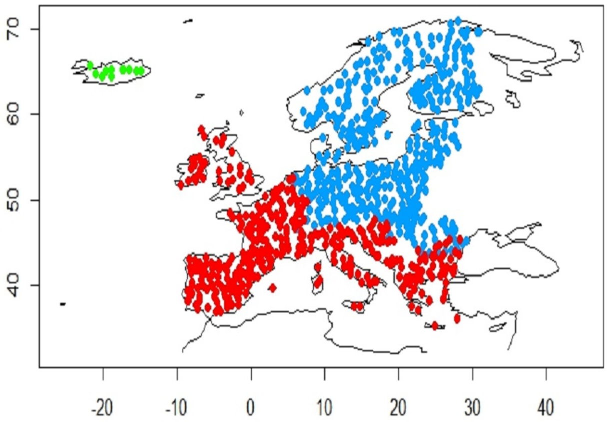

Figure 5.5: River basins in clusters with increasing losses in are depicted red and in clusters with decreasing losses in blue

	.9	.925	.95	.975	.99	.999	0.9999
1							
2020	$1.05 * 10^9$	$1.21 * 10^9$	$1.41 * 10^9$	$1.85 * 10^9$	$2.33 * 10^9$	$3.30 * 10^9$	$4.27 * 10^9$
2085	$4.63 * 10^8$	$6.41 * 10^8$	$8.97 * 10^8$	$1.29 * 10^9$	$1.84 * 10^9$	$2.87 * 10^9$	$3.76 * 10^9$
3							
2020	$3.60 * 10^8$	$6.34 * 10^8$	$1.13 * 10^9$	$2.49 * 10^9$	$5.40 * 10^9$	$1.21 * 10^{10}$	$1.69 * 10^{10}$
2085	$3.5 * 10^8$	$6.3 * 10^8$	$1.1 * 10^9$	$2.5 * 10^9$	$5.2 * 10^9$	$1.2 * 10^{10}$	$1.6 * 10^{10}$
4							
2020	$4.24 * 10^8$	$6.17 * 10^8$	$8.19 * 10^8$	$1.40 * 10^9$	$2.13 * 10^9$	$4.19 * 10^9$	$5.67 * 10^9$
2085	$3.69 * 10^8$	$6.22 * 10^8$	$7.84 * 10^8$	$1.34 * 10^9$	$2.05 * 10^9$	$4.04 * 10^9$	$5.53 * 10^9$
5							
2020	$2.34 * 10^8$	$3.53 * 10^8$	$5.74 * 10^8$	$1.21 * 10^9$	$2.32 * 10^9$	$4.85 * 10^9$	$6.00 * 10^9$
2085	$1.87 * 10^8$	$2.81 * 10^8$	$4.69 * 10^8$	$1.03 * 10^9$	$1.99 * 10^9$	$4.20 * 10^9$	$5.17 * 10^9$
6							
2020	$4.41 * 10^8$	$1.09 * 10^9$	$1.75 * 10^9$	$7.57 * 10^9$	$1.31 * 10^{10}$	$2.60 * 10^{10}$	$3.70 * 10^{10}$
2085	$4.90 * 10^8$	$1.20 * 10^9$	$1.94 * 10^9$	$8.57 * 10^9$	$1.43 * 10^{10}$	$2.75 * 10^{10}$	$4.06 * 10^{10}$
7							
2020	$1.77 * 10^8$	$4.61 * 10^8$	$9.72 * 10^8$	$2.42 * 10^9$	$5.82 * 10^9$	$1.39 * 10^{10}$	$2.07 * 10^{10}$
2085	$1.62 * 10^8$	$5.26 * 10^8$	$9.87 * 10^8$	$2.37 * 10^9$	$5.53 * 10^9$	$1.31 * 10^{10}$	$1.95 * 10^{10}$
8							
2020	$1.23 * 10^9$	$1.90 * 10^9$	$3.26 * 10^9$	$6.11 * 10^9$	$1.19 * 10^{10}$	$2.44 * 10^{10}$	$3.62 * 10^{10}$
2085	$1.05 * 10^9$	$1.58 * 10^9$	$2.88 * 10^9$	$5.36 * 10^9$	$1.03 * 10^{10}$	$2.16 * 10^{10}$	$3.20 * 10^{10}$
9							
2020	$8.37 * 10^8$	$1.97 * 10^9$	$3.09 * 10^9$	$7.50 * 10^9$	$1.64 * 10^{10}$	$4.04 * 10^{10}$	$5.39 * 10^{10}$
2085	$8.45 * 10^8$	$1.97 * 10^9$	$3.16 * 10^9$	$7.61 * 10^9$	$1.65 * 10^{10}$	$4.05 * 10^{10}$	$5.44 * 10^{10}$
10							
2020	$5.76 * 10^8$	$9.56 * 10^8$	$1.60 * 10^9$	$4.04 * 10^9$	$7.90 * 10^9$	$1.76 * 10^{10}$	$2.34 * 10^{10}$
2085	$6.12 * 10^8$	$9.62 * 10^8$	$1.59 * 10^9$	$4.01 * 10^9$	$7.59 * 10^9$	$1.69 * 10^{10}$	$2.27 * 10^{10}$
11							
2020	$1.60 * 10^7$	$1.22 * 10^8$	$5.09 * 10^8$	$2.59 * 10^9$	$9.86 * 10^9$	$3.05 * 10^{10}$	$3.96 * 10^{10}$
2085	$1.58 * 10^7$	$1.30 * 10^8$	$5.18 * 10^8$	$2.67 * 10^9$	$9.71 * 10^9$	$2.96 * 10^{10}$	$3.87 * 10^{10}$
12							
2020	$4.93 * 10^8$	$7.90 * 10^8$	$1.17 * 10^9$	$2.05 * 10^9$	$4.50 * 10^9$	$1.84 * 10^{10}$	$2.20 * 10^{10}$
2085	$5.31 * 10^8$	$8.01 * 10^8$	$1.29 * 10^9$	$2.11 * 10^9$	$5.27 * 10^9$	$1.81 * 10^{10}$	$2.24 * 10^{10}$
13							
2020	$3.18 * 10^8$	$4.59 * 10^8$	$6.31 * 10^8$	$1.12 * 10^9$	$2.02 * 10^9$	$4.20 * 10^9$	$5.28 * 10^9$
2085	$3.17 * 10^8$	$4.19 * 10^8$	$5.94 * 10^8$	$1.12 * 10^9$	$2.04 * 10^9$	$4.30 * 10^9$	$5.33 * 10^9$
14							
2020	$1.92 * 10^6$	$2.11 * 10^7$	$1.63 * 10^8$	$3.06 * 10^9$	$6.50 * 10^9$	$1.62 * 10^{10}$	$2.30 * 10^{10}$
2085	$2.25 * 10^6$	$2.27 * 10^7$	$1.78 * 10^8$	$3.64 * 10^9$	$7.73 * 10^9$	$1.86 * 10^{10}$	$2.69 * 10^{10}$
15							
2020	$1.02 * 10^9$	$1.53 * 10^9$	$2.05 * 10^9$	$3.60 * 10^9$	$5.44 * 10^9$	$9.84 * 10^9$	$1.42 * 10^{10}$
2085	$1.02 * 10^9$	$1.57 * 10^9$	$2.20 * 10^9$	$4.03 * 10^9$	$6.12 * 10^9$	$1.09 * 10^{10}$	$1.58 * 10^{10}$
16							
2020	$3.92 * 10^8$	$5.76 * 10^9$	$9.24 * 10^8$	$1.69 * 10^9$	$3.20 * 10^9$	$6.10 * 10^9$	$7.14 * 10^9$
2085	$9.95 * 10^8$	$5.82 * 10^9$	$9.59 * 10^8$	$1.71 * 10^9$	$3.20 * 10^9$	$6.11 * 10^9$	$7.20 * 10^9$

Table 5.4: Value at risk for the clusters in 2020 and 2085.

	.9	.925	.95	.975	.99	.999	.9999
1							
2020	$1.04 * 10^9$	$1.22 * 10^9$	$1.46 * 10^9$	$1.85 * 10^9$	$2.33 * 10^9$	$3.30 * 10^9$	$4.27 * 10^9$
2085	$1.06 * 10^9$	$1.23 * 10^9$	$1.46 * 10^9$	$1.84 * 10^9$	$2.31 * 10^9$	$3.26 * 10^9$	$4.13 * 10^9$
3							
2020	$2.15 * 10^9$	$2.70 * 10^9$	$3.61 * 10^9$	$5.54 * 10^9$	$8.18 * 10^9$	$1.42 * 10^{10}$	$1.81 * 10^{10}$
2085	$2.11 * 10^9$	$2.66 * 10^9$	$3.55 * 10^9$	$5.41 * 10^9$	$7.90 * 10^9$	$1.37 * 10^{10}$	$1.75 * 10^{10}$
4							
2020	$1.12 * 10^9$	$1.32 * 10^9$	$1.63 * 10^9$	$2.22 * 10^9$	$3.02 * 10^9$	$4.86 * 10^9$	$6.21 * 10^9$
2085	$1.08 * 10^9$	$1.28 * 10^9$	$1.56 * 10^9$	$2.13 * 10^9$	$2.91 * 10^9$	$4.71 * 10^9$	$6.05 * 10^9$
5							
2020	$9.87 * 10^8$	$1.22 * 10^9$	$1.60 * 10^9$	$2.37 * 10^9$	$3.43 * 10^9$	$5.42 * 10^9$	$6.16 * 10^9$
2085	$8.33 * 10^8$	$1.03 * 10^9$	$1.36 * 10^9$	$2.04 * 10^9$	$2.96 * 10^9$	$4.68 * 10^9$	$5.30 * 10^9$
6							
2020	$4.84 * 10^9$	$6.21 * 10^9$	$8.63 * 10^9$	$1.37 * 10^{10}$	$1.83 * 10^{10}$	$3.06 * 10^{10}$	$4.25 * 10^{10}$
2085	$5.31 * 10^9$	$6.82 * 10^9$	$9.48 * 10^9$	$1.51 * 10^{10}$	$2.00 * 10^{10}$	$3.33 * 10^{10}$	$4.61 * 10^{10}$
7							
2020	$2.14 * 10^9$	$2.74 * 10^9$	$3.77 * 10^9$	$6.02 * 10^9$	$9.27 * 10^9$	$1.75 * 10^{10}$	$2.27 * 10^{10}$
2085	$2.05 * 10^9$	$2.63 * 10^9$	$3.61 * 10^9$	$5.73 * 10^9$	$8.80 * 10^9$	$1.65 * 10^{10}$	$2.15 * 10^{10}$
8							
2020	$5.13 * 10^9$	$6.33 * 10^9$	$8.24 * 10^9$	$1.20 * 10^{10}$	$1.77 * 10^{10}$	$2.93 * 10^{10}$	$4.01 * 10^{10}$
2085	$4.48 * 10^9$	$5.55 * 10^9$	$7.24 * 10^9$	$1.06 * 10^{10}$	$1.55 * 10^{10}$	$2.59 * 10^{10}$	$3.53 * 10^{10}$
9							
2020	$6.43 * 10^9$	$8.10 * 10^9$	$1.09 * 10^{10}$	$1.71 * 10^{10}$	$2.66 * 10^{10}$	$4.72 * 10^{10}$	$5.76 * 10^{10}$
2085	$6.52 * 10^9$	$8.19 * 10^9$	$1.11 * 10^{10}$	$1.73 * 10^{10}$	$2.67 * 10^{10}$	$4.74 * 10^{10}$	$5.86 * 10^{10}$
10							
2020	$3.18 * 10^9$	$3.99 * 10^9$	$5.37 * 10^9$	$8.09 * 10^9$	$1.20 * 10^{10}$	$2.03 * 10^{10}$	$2.57 * 10^{10}$
2085	$3.11 * 10^9$	$3.89 * 10^9$	$5.23 * 10^9$	$7.81 * 10^9$	$1.15 * 10^{10}$	$1.95 * 10^{10}$	$2.48 * 10^{10}$
11							
2020	$3.05 * 10^9$	$4.04 * 10^9$	$5.94 * 10^9$	$1.04 * 10^{10}$	$1.86 * 10^{10}$	$3.47 * 10^{10}$	$4.09 * 10^{10}$
2085	$3.02 * 10^9$	$4.00 * 10^9$	$5.87 * 10^9$	$1.02 * 10^{10}$	$1.81 * 10^{10}$	$3.38 * 10^{10}$	$3.99 * 10^{10}$
12							
2020	$2.35 * 10^9$	$2.92 * 10^9$	$3.89 * 10^9$	$6.21 * 10^9$	$1.13 * 10^{10}$	$1.97 * 10^{10}$	$2.30 * 10^{10}$
2085	$2.45 * 10^9$	$3.05 * 10^9$	$4.05 * 10^9$	$6.42 * 10^9$	$1.14 * 10^{10}$	$1.96 * 10^{10}$	$2.34 * 10^{10}$
13							
2020	$9.61 * 10^8$	$1.15 * 10^9$	$1.46 * 10^9$	$2.09 * 10^9$	$3.03 * 10^9$	$4.71 * 10^9$	$5.52 * 10^9$
2085	$9.52 * 10^8$	$1.15 * 10^9$	$1.47 * 10^9$	$2.12 * 10^9$	$3.08 * 10^9$	$4.81 * 10^9$	$5.62 * 10^9$
14							
2020	$2.11 * 10^9$	$2.81 * 10^9$	$4.20 * 10^9$	$7.08 * 10^9$	$1.03 * 10^{10}$	$1.88 * 10^{10}$	$2.53 * 10^{10}$
2085	$2.45 * 10^9$	$3.26 * 10^9$	$4.88 * 10^9$	$8.25 * 10^9$	$1.20 * 10^{10}$	$2.19 * 10^{10}$	$2.92 * 10^{10}$
15							
2020	$2.88 * 10^9$	$3.42 * 10^9$	$4.23 * 10^9$	$5.64 * 10^9$	$7.44 * 10^9$	$1.19 * 10^{10}$	$1.58 * 10^{10}$
2085	$3.10 * 10^9$	$3.71 * 10^9$	$4.62 * 10^9$	$6.22 * 10^9$	$8.16 * 10^9$	$1.32 * 10^{10}$	$1.75 * 10^{10}$
16							
2020	$1.43 * 10^9$	$1.75 * 10^9$	$2.25 * 10^9$	$3.22 * 10^9$	$4.65 * 10^9$	$6.58 * 10^9$	$7.53 * 10^9$
2085	$1.44 * 10^9$	$1.77 * 10^9$	$2.27 * 10^9$	$3.24 * 10^9$	$4.65 * 10^9$	$6.61 * 10^9$	$7.57 * 10^9$

Table 5.5: Expected shortfall for the clusters in 2020 and 2085

Chapter 6

Conclusions and outlook

The first conclusion we draw from this work is that the dependence between the behaviour of rivers in geographic proximity prevent a sufficient hedging of the risk originating from flood events unless sufficiently large areas are considered. Therefore cooperation on the national, and especially for smaller nations, on the international level is needed to provide sufficient reserves without too expensive costs.

Our results show that in the most affected areas the expected damages will increase significantly. This suggests that for the strongly affected areas it becomes necessary to increase flood protection to counteract the increased risk caused by the increased precipitation triggered by climate change.

We hope that this work does provide insight why the dependency between random variables cannot be ignored when one deals with risk presented by natural hazards and that the findings of this work can be used to improve precautions and reserves for flood hazards in Europe.

We find that the work and results could be improved by including further observations and inclusion of the so far missing forecasts for some of the loss areas. An increased number of observations would allow to increase the fit of the models in this work and would improve the power of the statistical test we use throughout the work.

In addition, with more observation controlling for seasonal effects would become viable which would allow a more nuanced separation of the individual rivers into clusters and a better forecast for individual smaller regions.

Appendix A

Additional Tables

copula	generator function	parameter range
Product	$-ln(t)$	
Clayton	$\frac{1}{\theta}(t^{-\theta} - 1)$	$[-1, \inf) \setminus 0$
Joe	$-ln(1 - (1 - t)^{\theta})$	$[0, \inf)$
Gumbel	$(-ln(t))^{\theta}$	$[0, \inf)$
Frank	$ln(e^{-\theta} - 1) - ln(e^{-t\theta} - 1)$	$(-\inf, \inf) \setminus 0$

Table A.1: Generator functions and parameter ranges for Archimedean copulas

copula	Kendall's τ	λ_l	λ_u
Clayton	$\theta/(\theta + 2)$	$2^{-1/\theta}$	0
Joe	$1 + \frac{4}{\theta^2} \int_0^1 x \cdot ln(x)(1 - x)^{\frac{2(1-\theta)}{\theta}} dx$	0	$2 - 2^{-1/\theta}$
Gumbel	$1 - 1/\theta$	0	$2 - 2^{-1/\theta}$
Frank	$1 - \frac{4}{\theta^2} + 4D_1(\theta)/\theta$	0	0
Gaussian	$\frac{2}{\pi} arcsin(\rho)$	0	0

Table A.2: Kendall's τ, λ_l and λ_u for the copulas used in this work. $D_1(\theta) = \int_0^\theta \frac{x/\theta}{e^x - 1} dx$.

$$\begin{pmatrix} 1 & & & & & \\ 6 & 4 & & & & \\ 3 & 6 & 5 & & & \\ \underline{4} & 3 & \underline{6} & 2 & & \\ 5 & 2 & 3 & 6 & 3 & \\ 2 & 5 & 2 & 3 & 6 & 6 \end{pmatrix}$$

Table A.3: An alternative array representation for the R.Vine in figure 2.7

Cluster	countries
1	Norway, Sweden, Finland
2	Iceland
3	Finland
4	Denmark,Sweden
5	Estonia, Latvia, Lithuania
6	Ireland,UK,Belgium,Netherlands
7	Germany
8	Poland,Czech R.,Slovakia
9	France
10	Romania, Hungary
11	Switzerland,Austria
12	Croatia,Bosnia,Albania, Greece,Southern Italy
13	Bulgaria,Romania
14	Northern Italy
15	Northern Spain, Southern France, Italy
16	Spain, Portugal

Table A.4: The allocation of clusters to European countries.

Appendix B

Proofs

The information distance is a metric.

A metric must fulfil four requirements:

- $d_{\hat{\tau}}(X,Y) \geq 0$: Kendall's τ only takes values in the interval $(-1,1)$, positivity of $d_{\hat{\tau}}$ is ensured by construction.
- $d_{\hat{\tau}}(X,Y) = 0 \Leftrightarrow X \sim Y$.

$$d_{\hat{\tau}}(X,Y) = 0 \Leftrightarrow \tau_{X,Y} \in \{-1,1\} \Leftrightarrow X = Y \vee X = \overleftarrow{Y} \Leftrightarrow X \sim Y$$

- $d_{\hat{\tau}}(X,Y) = d_{\hat{\tau}}(Y,X)$ this is true since $\tau_{X,Y} = \tau_{Y,X}$
- $d_{\hat{\tau}}(X,Z) \leq d_{\hat{\tau}}(X,Y) + d_{\hat{\tau}}(Y,Z)$

$$\begin{aligned}
&d_{\hat{\tau}}(X,Z) = \\
&min\{d_\tau(X,Z), d_\tau(X,\overleftarrow{Z}), d_\tau(X,\overleftarrow{Z}), d_\tau(X,Z)\} \leq \\
&min\{d_\tau(X,Y) + d_\tau(Y,Z), d_\tau(X,Y) + d_\tau(Y,\overleftarrow{Z}), \\
&d_\tau(X,\overleftarrow{Y}) + d_\tau(\overleftarrow{Y},\overleftarrow{Z}), d_\tau(X,\overleftarrow{Y}) + d_\tau(\overleftarrow{Y},Z)\} = \\
&min\{d_\tau(X,Y), d_\tau(X,\overleftarrow{Y})\} + min\{d_\tau(Y,Z), d_\tau(Y,\overleftarrow{Z})\} = \\
&d_{\hat{\tau}}(X,Y) + d_{\hat{\tau}}(Y,Z)
\end{aligned}$$

□

Sklar's theorem for continuous margins.

Let H be the joined distribution function of two random variables X, Y with continuous marginal distributions $F(x), G(y)$. Let $\forall u \in [0,1] \quad F^{-1}(u) := inf\{x : F(x) \geq u\}$, $\forall v \in [0,1] \quad G^{-1}(v) := inf\{y : G(y) \geq v\}$. Then

$$C(u,v) := H(F^{-1}(u), G^{-1}(v))$$

. By construction C full-fills all requirements for a copula.
The uniqueness of C also follows from the continuity of the marginal distributions. Assume there exist two different copulas $C, \tilde{C}$ for all $u \in [0,1]$ there exists an x such that $F(x) = u \Leftrightarrow F^{-1}(u) \neq \emptyset$ the same is true for v this implies that $\forall (u,v) \in [0,1]^2 \quad C(u,v) = \tilde{C}(u,v) \Rightarrow C = \tilde{C}$. □

Bibliography

[1] E.C. Brechmann;C. Czado; K. Aas. Truncated regular vines in high dimensions with application to financial data. *Canadian Journal of Statistics*, 40:68–85, 2012.

[2] F. N. Fritsch; R. E. Carlson. Monotone piecewise cubic interpolation. *SIAM J.*, 17:238–246, 1980.

[3] T. Bedford; R.M. Cooke. Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence*, 32:245–268, 2001.

[4] D. Müller;C. Czado. Representing sparse gaussian dags as sparse r-vines allowing for non-gaussian dependence, 2016.

[5] J. C. Dunn. Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4(1):95–104, 1974.

[6] A.J. McNeil;R. Frey;P. Embrechts. *Quantitative Risk Management: Concepts, Techniques, and Tools.* Princeton University Press, 2005.

[7] B. Jongman et al. Increasing stress on disaster-risk finance due to large floods. *Nature Climate Change*, 4:264–268, 2014.

[8] T. Calinski; J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974.

[9] C. Czado; S. Jeske; M. Hofmann. Selection strategies for regular vine copulae. *Journal de la Société Française de Statistique*, 154:174–191, 2013.

[10] H. Joe. Families of m-variate distributions with given margins and m(m-1)/2 bivariate pependence parameters. *Distributions with Fixed Marginals and Related Topics*, 28, 1996.

[11] J.H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244, 1963.

[12] M.G. Kendall. A new measure of rank correlation. *Biometrika*, 30:81–93, 1938.

[13] D. Kurowicka. Optimal truncation of vines. *VinesHandbook*, pages 233–247, 2011.

[14] J.F. Dissmann; C.Czado; D. Kurowicka. Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics Data Analysis*, 59:52–69, 2013.

[15] O. Morales-Nápoles; R.M. Cooke; D. Kurowicka. About the number of vines and regular vines on n nodes, 2009.

[16] R.B. Nelson. *An Introduction to copulas.* Springer, 1999.

[17] J. M. Van Der Knijff; J. Younis ; A. P. J. De Roo. Lisflood: a gis-based distributed model for river basin scale water balance and flood simulation. *International Journal of Geographical Information Science*, 24:2:189–212, 2010.

[18] A. Sklar. *Fonctions de Répartition À N Dimensions Et Leurs Marges.* Université Paris 8, 1959.

[19] Q. H. Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57:307–333, 1989.

[20] H.J. Beilharz; B. Rauch; C. Wallner. Insurance aids economic recovery. *Topics Geo*, pages 13–14, 2014.

[21] G. N. Lance; W. T. Williams. A general theory of classificatory sorting strategies1. hierarchical systems. *The Computer Journal*, 9(4):373–380, 1967.

Abstract

In the recent years Europe was hit several times by large-scale flooding events. This makes it rather clear that in order to estimate damages and resulting claims to insurance companies it is important not only to analyse each body of water independently from the rest, but to also examine the joint distribution of the water discharges in the individual basins.

In order to model the joint distribution we use R-Vine-copulas. First we use historic data on the monthly peak water discharge of the river basins to calculate the R-Vine-copulas. Then we use these R-Vines combined with individual forecasts for the rivers to sample from the joint distribution and build the cumulative distribution function for the sum of total losses.

From these cumulative distribution functions we calculate some risk measures like the value at risk and the expected shortfall in order to estimate the impact of climate change on the frequency and magnitude of damages caused by floods in Europe.

Abstract

In den letzten Jahren gab mehrere großflächige Überflutungen in Europa. Um Flutschäden und die daraus entstehenden Forderungen an Versicherungen abschätzen zu können ist es daher wichtig nicht nur die einzelnen Flüsse zu betrachten sondern auch die gemeinschaftliche Verteilung der Durchflussmenge der Flüsse zu berücksichtigen.

Zur Modellierung der gemeinschaftlichen Verteilung verwenden wir R-Vine-Copulas. Zuerst benutzen wir Daten für die monatliche Höchstdurchflussmenge der einzelnen Flüsse in der Vergangenheit um die R-Vine-Copulas zu berechnen Danach benutzen diese R-Vine zusammen mit Vorhersagen zu den einzelnen Flüssen um Stichproben aus der gemeinschaftlichen Verteilung zu ziehen und die Verteilungsfunktion für die Summe der Schäden zu bilden.

Mit Hilfe dieser Verteilungsfunktionen berechnen wir einige Risikomaße wie Value at Risk und Expected Shortfall um die Auswirkungen des Klimawandels auf Vorkommen und Höhe von Flutschäden in Europa abzuschätzen.